Understanding and Improving the Identification of Somatic Variants

Vinaya Vijayan

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Genetics, Bioinformatics and Computational Biology

Chair: Liqing Zhang

Christopher Franck

Lenwood S. Heath

Xiaowei Wu

August 12th, 2016

Blacksburg, VA, USA

Keywords: Somatic variants, Somatic variant callers, Somatic point mutations, Short tandem repeat variation, Lung squamous cell carcinoma

Understanding and Improving the Identification of Somatic Variants

Vinaya Vijayan

ABSTRACT

It is important to understand the entire spectrum of somatic variants to gain more insight into mutations that occur in different cancers for development of better diagnostic, prognostic and therapeutic tools. This thesis outlines our work in understanding somatic variant calling, improving the identification of somatic variants from whole genome and whole exome platforms and identification of biomarkers for lung cancer.

Integrating somatic variants from whole genome and whole exome platforms poses a challenge as variants identified in the exonic regions of the whole genome platform may not be identified on the whole exome platform and vice-versa. Taking a simple union or intersection of the somatic variants from both platforms would lead to inclusion of many false positives (through union) and exclusion of many true variants (through intersection). We develop the first framework to improve the identification of somatic variants on whole genome and exome platforms using a machine learning approach by combining the results from two popular somatic variant callers. Testing on simulated and real data sets shows that our framework identifies variants more accurately than using only one somatic variant caller or using variants from only one platform.

Short tandem repeats (STRs) are repetitive units of 2-6 nucleotides. STRs make up approximately 1% of the human genome and have been traditionally used as genetic markers in population studies. We conduct a series of *in silico* analyses using the exome data of 32 individuals with lung cancer to identify 103 STRs that could potentially serve as cancer diagnostic markers and 624 STRs that could potentially serve as cancer predisposition markers.

Overall these studies improve the accuracy in identification of somatic variants and highlight the association of STRs to lung cancer.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Liqing Zhang, for letting me join her lab while continuing to work on my previous projects and mentoring me through those projects. She has been a terrific mentor and I will eternally be grateful to her for sharing her scientific wisdom every day and chocolates during our lab meetings.

I also thank my committee members, Lenwood Heath, Christopher Franck, Xiaowei Wu, for believing in me and for guiding me through tough times. They also provided consistent feedback and valuable insights in my projects.

My previous mentors, David Mittelman, Pawel Michalak, Andy Pereira, whose encouragement, support, critique, different styles of working, and different perspectives in science have enriched my experience and helped me grow as a scientist.

My initial mentors at National Chemical Laboratory, India, Chetan Gadgil and Mugdha Gadgil, who sparked my interest in research and encouraged me to pursue further studies.

My labmates over the years, Arjun and Zalman, who guided me in the right direction through the initial stages of my PhD. Endless discussions with Robin, Gareth, Mohammad, Tithi, Hong and Gustavo about science and life not only helped me understand different aspects of science but also helped me grow as a person.

Friends I have made in Blacksburg, Abhijit, Siddhesh, Deepanshu, Raj, Revathy, Sarthak, Saloni, Primal, Nistha and Khushboo who have helped make Blacksburg home away from home. Faizan, for all the weekend phone calls, for not letting distance or the fact that we have not met in around 8 years ever get in the way of an amazing friendship. Shruti, for patiently listening to my endless rants and for being an incredible source of answers for my questions on the personal and professional front.

My parents, Vimala and N.K. Vijayan for encouraging and backing up all our dreams through their love, hard work and perseverance. All I can say is thank you and hope those words convey my gratitude towards them. My brother, Vikas, for providing the comic relief in my life regardless of whether it was required. My husband, Nikhil, for his unwavering support and sharing this roller coaster ride of my PhD and life!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter One

Literature Review

1.1 Human Genome Project

The completion of the Human Genome Project (HGP) in 2003 marks one of the greatest milestones in genetics. The HGP consisted of an international scientific consortium whose aim was to determine the order of chemical bases called deoxyribonucleic acid (DNA) that make up the human genome. The germination of the project was with different discussions that began in 1984. The project finally went underway in 1990 with collaborative efforts from Department of Energy (DOE), National Institute of Health (NIH), and other international agencies [1]. A rough draft of the human genome consisting of 3 billion base pairs almost 94% complete was released in 2000 [1]. The HGP was finally declared complete in April 2003 [2]. The sequencing and assembling of the first human genome took 13 years and was done at a cost of nearly $3 billion.

Different techniques that enabled the sequencing of the first draft of the human genome include cloning systems such as cosmids, yeast artificial chromosomes (YAC), and bacterial artificial chromosomes (BAC), and sequencing techniques such as capillary gel electrophoresis, four-color fluorescence-based sequence detection, dye labeled terminators and cycle sequencing [1]. The draft genome was assembled with the help of software packages such as PHRED, PHRAP, and GigAssembler [1].

1.2 Next Generation Sequencing Platforms

The advent of high throughput sequencing around 2004 has helped reduce the cost of sequencing exponentially and the time in which DNA of an organism can be sequenced [3].

Roche/454 FLX Pyrosequencer was the first next-generation sequencer, which was introduced in 2004 and uses a sequencing technology called pyrosequencing [4]. The SOLiD platform, introduced in 2006, uses an emulsion PCR approach to sequence DNA [4]. Illumina, still the most widely used platform for sequencing DNA, conducts sequencing by a synthesis approach [4] that relies on the detection of fluorescence generated by the addition of fluorescent labeled reversible-terminator nucleotides to the existing DNA strand. Ion Torrent uses semiconductor sequencing to detect the pH change during the release of a hydrogen ion when a nucleotide gets added to the DNA strand [5]. PacBio utilizes single molecule real time (SMRT) technology to detect fluorescent light pulses generated during incorporation of a nucleotide to the DNA strand at the bottom of a zero mode waveguide (ZMW) chamber [6]. The read length has seen an increase from the initial days of 25-35 bp reads [4] to PacBio now generating read lengths of over 10kb [7]. The release of the HiSeq X Ten by Illumina also brings us close to the dream of producing $1000 genome, thus making sequencing much more affordable [8]. This decrease in cost for sequencing is even faster than Moore's law, as shown by the National Human Genome Research Institute [9]. Once genomes or exomes have been sequenced, the next downstream process is to identify variants.

1.3 Germline and Somatic Variants

There are two types of variants in terms of sequencing in different tissues. Mutations that occur in germline tissues and can be inherited by the progeny are called germline mutations. Mutations that occur in somatic tissues and are not inherited from the previous generation are somatic mutations. These mutations can be point mutations, indels, structural mutations, copy

number variations, and short tandem repeat variations. A point mutation is the change of a single nucleotide. An indel is a segment of DNA that gets inserted into or deleted from the genome. Structural variants are rearrangements of segments of DNA, larger than 1kb regions, that include inversions, translocations, insertions, and deletions and that can be inter or intra-chromosomal [10]. Copy number variation is a sub-class of structural variants that show variation in the copy number of a segment of DNA. Short tandem repeats (STR) are motifs of 2-6 nucleotides that get repeated multiple times on the chromosome. Approximately 1% of the human genome is composed of STRs. The number of times an STR gets repeated in a particular region can vary in different individuals. This variation could also be responsible for differential gene expression [11]. Identification of these variants has been helpful in our understanding of human variations.

1.4 Uses of Sequencing

An important discovery of HGP was that the number of genes present in a human is approximately 30,000 rather than the estimated number of 40,000-100,000 initially thought [1]. The HGP and subsequent rise in high throughput sequencing has led to the discovery of many disease causing genes [12]. This also owes to the fact that the discovery of genetic aberrations is now less time consuming as compared to the time when the human genome was not available. Sequencing has shed light on rare diseases [13], helped find alternate pathways that affect the same disease [14], and has led to the development of a new field called precision medicine [15].

High throughput sequencing of tumor and normal samples has given us new insight into cancer. Sequencing the human genome has also led to the discovery of oncogenes, and ongoing

projects aim to identify all the genetic abnormalities seen in 50 major types of cancer [16]. High throughput sequencing of genomes, exomes, and gene expression data of individuals with cancer has been done to improve diagnosis and prognosis, and to garner information on different stages of cancer in order to identify different therapeutic agents [17]. There are different kinds of genetic tests that help determine the predisposition of a person to any disease and prenatal tests to find if an unborn child may suffer from a genetic disease. The human genome sequence has helped find alternate causes for known diseases [18]. Pharmacogenomics is a new emerging field where the design of drugs and its effects on genetic variation in individuals is studied [19]. The sequencing of the human genome has also led to insight into evolution [20] and advances in forensic science [21]. The development of these fields has been through concerted collaborations of various institutions from different countries.

1.5 Next Generation Sequencing Collaborations

Many academic and industrial next generation sequencing collaborations over the past few years have led to improved understanding of human variations with respect to ancestry, genealogy, and diseases. The UCSC Genome Browser, Encyclopedia of DNA Elements (ENCODE) and National Center for Biotechnology Information (NCBI) host various information on human genomes [22]. The 1000 Genomes project catalogs over 3,000 human genomes, and provides insight into the variations harbored in different human populations [23]. Consortiums, such as the Exome Aggregation Consortium [24], CARDIoGRAMplusC4D Consortium [25], Diabetes Genetics Initiative [26], and Autism Consortiums [27] are some of the many consortiums that aim to increase our understanding of rare diseases, cardiovascular diseases, type

2 diabetes, and autism respectively. Variations in different genes have been catalogued in HapMap and dbSNP [28]. Databases such as DisGeNet and Domain Mapping of Disease Mutations (DM^2) record the association of various mutations with diseases [29]. Functional annotation tools such as KEGG and DAVID help understand which diseases or cellular processes might be affected if a particular set of genes are affected [30]. The Cancer Genome Atlas (TCGA) is built by ICGC to increase our comprehension of cancer [31]. TCGA catalogs whole genome, whole exome, and RNASeq data and records all the variants found in the Catalog of Somatic Mutations in Cancer (COSMIC) [32]. Industries such as 23andMe have also helped catalogue DNA sequence data from consenting individuals to improve our understanding of diseases such as Parkinson's disease, Alzheimers, and Prion disease [33].

1.6 Mapping Tools

Alignment tools have evolved according to different technological advances in sequencing and the increasing amounts of high throughput sequencing data [34]. There are many different kinds of alignment tools to map DNA, RNA, short RNA and bisulfite data. As mentioned before, read lengths have increased from the original 25-35 bps to lengths of more than 10kb. Protocols to develop these reads have evolved from single-end reads to paired-end reads. Aligners have had to adapt to all these technological advances. Fonseca et al. [34] and wiki pages [35] have a detailed survey of the aligners that would suit different purposes. Various alignment tools such as BWA, BWA-MEM, BWA-SW, Bowtie2, Novoalign, Stampy, SHRiMP, and TMAP have been built over the years to map reads to a reference genome. BWA, BWA-MEM, and BWA-SW were built by the same group but for different read lengths. BWA, BWA-

MEM, BWA-SW, and Bowtie use the Burrows Wheeler Transform and associated data structures to map reads through long inexact string matching [36]. Novoalign uses Needleman-Wunsch to optimally align reads globally with affine gap penalties [37]. TMAP was built specifically for Ion Torrent data taking into consideration the fact that Ion Torrent produces more indels than substitution errors which is typical for Illumina [38]. SHRiMP2 indexes the genome and optimally aligns reads using the Smith-Waterman algorithm [39]. The earlier version of SHRiMP indexed the reads instead of the genome [40]. Stampy uses a hash table data structure and Smith-Waterman and a hybrid error model to map the reads [41]. Out of these, BWA and Bowtie2 are the most popular tools used by the high throughput sequencing community. The 1000 Genomes Project and TCGA usually map their data using BWA.

1.7 Germline Variant Calling Tools

Once short reads are aligned to the reference genome, the next process is to determine the variants that lie in the sequenced individuals, which can be helpful for clinical and research purposes. Just like the increasing number of aligners, variant calling tools have also been on the rise. There are some variant callers custom made for a particular sequencing platform or aligner [42], but most of the variant callers can be used for any high throughput sequencing data regardless of the platform or alignment tool [43]. Germline variant calling tools include GATK, FreeBayes, Atlas, SAMtools, and Isaac, among others. Isaac and TMAP were built by Illumina and Ion Torrent respectively to identify germline variants from data generated on their own sequencing platforms [44]. GATK HaplotypeCaller was built by the Broad Institute to identify germline point variants and indels [45]. HaplotypeCaller does local de novo assembly of regions

where variants are found to accurately identify germline variants. SAMtools and GATK use Bayesian models to distinguish germline variants from sequencing errors [46]. Atlas2 uses logistic regression models trained on whole exome capture sequencing data to identify germline variants [47]. All of the above mentioned germline variant callers identify point mutations and indels.

1.8 Somatic Variant Calling Tools

The technological advances and high resolution of sequencing has made it possible to identify somatic variants. Identifying these variants is the first step in the process of studying the mutated genes or pathways significant in a particular cancer. The development of tools identifying somatic variants has seen a rise in the past 3-4 years. The tools typically compare a tumor sample with a matched normal sample to identify somatic variants [48], but there are also tools that could identify somatic variants using only tumor samples [49]. Most of the former aim for a high precision whereas the latter aim for a high sensitivity. The somatic variant callers make use of different parameters, such as read depth, strand bias, quality score of the reads, and number of variants identified in a specific window, to determine somatic point mutations, somatic indels, copy number variants, or structural variants. MuTect and SomaticSniper use a Bayesian approach to identify somatic variants [50], while VarScan2 and VarDict employ Fisher's exact test to identify somatic variants [48]. VCMM uses a simple multinomial model to distinguish a potential candidate somatic variant from a sequencing error [49]. VarScan2 and VarDict are some of the few somatic indel callers, and there is much room for developing new somatic indel callers. Aneuploidy, polyploidy, and heterogeneity of cells are characteristic of

tumor samples. To detect these copy number alterations and regions with loss of heterozygosity, copy number variant callers for tumor samples have been developed, which include SegSeq [51], ReadDepth [52], BIC-seq [53], Patchwork [54], OncoSNP-SEQ [55], HMMcopy [56], and CONSERTING [57] for whole genome platforms [58]. ExomeCNV [59], VarScan2 [48], and HAPSEG/ABSOLUTE [60] are copy number variant callers for whole exome platforms. Control_FREEC [61] can call somatic copy number variants from whole genome and whole exome platforms. BreakDancer [62], Breakpointer [63], CLEVER [64], GASVPro [65], SVMerge [66], and MetaSV [67] are tools developed for identification of structural variants that include long indels spanning over 1kb, inter and intra-chromosomal translocations, and inversions [68].

1.9 Short Tandem Repeat Calling Tools

Calling short tandem repeats (STRs) accurately is challenging for a variety of reasons [69]. During PCR amplification or during bridge amplification carried out by the sequencer, stutter noise is created by DNA polymerase due to slippage events. Reads that do not encompass the entire STR region cannot be trusted as they could cause an error in identifying the number of times a motif repeats. There are two main tools, LobSTR [69] and RepeatSeq [70], developed to identify the number of times a motif repeats, which have also been used to identify the short tandem repeats in all the genomes of the 1000 Genomes Project [71]. LobSTR calls short tandem repeats in three steps: (i) sensing - identify the motif that repeats in the locus; (ii) alignment- realign the reads in the locus; and (iii) allelotyping - identify the number of times the motif repeats [69]. RepeatSeq uses a Bayesian model approach guided by an error model to

determine the STR allele [70]. Willems et al. have shown that LobSTR works better than RepeatSeq in determining STR alleles [71].

1.10 Research Plan

There are many sequencing platforms, alignment tools, variant callers and somatic variant callers for generating and analyzing exome and genome data. The previous bottleneck in genomics was data collection. However, due to the reduction in cost and time associated with sequencing genomes and exomes, the bottleneck has shifted to the analysis of high throughput sequencing data. To improve our understanding of which pipeline would work better to identify variants under different conditions, such as differing sequencing depths, allele fractions, mapping qualities, and base qualities, it was essential to review the current variant callers. The analysis and study of these variants is necessary from a clinical and research point of view. Chapter Two describes a comparison of the existing pipelines encompassing several popular mappers and somatic variant callers. Here, we design simulated data sets and define metrics to understand the performance of pipelines under various conditions on different platforms, i.e., whole genomes and whole exomes with the aim to use different pipelines for clinical and research purposes. The mappers used in our study (BWA, BWA-MEM, and Bowtie2) do not affect somatic variant calling as much as somatic variant calling. Among the somatic variant callers (Mutect, SomaticSniper, VarScan2, and VCMM), MuTect has the highest precision and VCMM has the highest sensitivity.

While reviewing different algorithms it was clear that the concordance between variants found in the whole exome and exonic regions of whole genomes is very low (<11%). This was also seen in the case of germline variants previously. Other studies also indicated that analysis of whole genomes to identify variants was more fruitful than identifying variants from whole exomes for germline variants. If a similar case arises, i.e., if variants are found during our whole genome analysis but not in the whole exome analysis and vice-versa we are unsure of which of the two platforms should be trusted. Chapter Three describes a framework for improving identification of somatic variants while using whole exome and whole genome platforms. This framework utilizes the outputs of Mutect and VCMM as features for the decision tree algorithm. The decision tree then uses a well-defined training set to accept or reject a given variant as somatic. Our framework, which has been tested on both simulated and real data, identifies variants more accurately than using only one somatic variant caller or using variants from only one platform (i.e., only whole genome or only whole exome).

Chapter Four describes *in silico* identification of short tandem repeat variations that might be important indicators of having lung cancer or cancer predisposition. Short tandem repeats have been used as genetic markers for a long time in paternity testing and forensic sciences and for over 40 Mendelian diseases such as Huntington's disease, neurodegenerative diseases, and muscular dystrophy. We hypothesize that short tandem repeats can also be used as genetic markers for lung cancer. Using the exome data of 32 individuals with lung cancer, we conducted a series of *in silico* analyses to identify 103 STRs that could potentially serve as cancer diagnostic markers and 624 STRs that could serve as cancer predisposition markers.

Chapter Two

Evaluation of pipelines detecting somatic point variants and analysis of factors affecting the detection

2.1 Abstract

Somatic variation is an important cause of cancer. The advent of next generation sequencing (NGS) has made possible the identification of cancer driving somatic variants in patients. Many short read mappers and somatic variant callers have been introduced for identifying somatic mutations in NGS data. With an increasing number of such computational tools made available, it becomes challenging for researchers to decide which ones to use, thus there is a need for comparing the existing pipelines to better guide the identification of somatic mutations. Towards this end, we compared the performance of entire somatic point mutation identification pipelines including three short read mapping algorithms (BWA, BWA-MEM, Bowtie2) and four somatic variant callers (MuTect, VarScan2, SomaticSniper, VCMM) on both simulated tumor data sets and real breast cancer data sets of exomes and genomes. Results show that any of the mappers BWA, BWA-MEM, or Bowtie2 can be used to detect somatic variants with similar results. VCMM is the best somatic variant caller for somatic point mutations that can be used if false positives are not an issue. MuTect is the most conservative somatic point mutation caller that can be used for the least number of false positives and a lesser number of false negatives compared to VCMM. The study provides guidelines to the users on the use of different mappers and somatic variant callers for the identification of somatic point variants according to the specific requirement of the users.

Keywords: Somatic variant calling, Next-generation sequencing, Software evaluation

2.2 Introduction

The advent of next generation sequencing (NGS) has helped reduce the cost of sequencing and the time in which a genome can be sequenced. The release of HiSeq X Ten by Illumina also brings us close to the dream of producing a $1000 genome, thus making sequencing much more affordable [8]. The first human genome was sequenced in a span of 13 years [72], while one can now be sequenced within a few days. Read lengths have increased from 25-35 bps [73] during the initial days to now, for example, PacBio generating read lengths of over 10kb [74]. Protocols to develop these reads have evolved from single-end reads to paired-end reads.

Alignment tools have evolved according to different technological advances in sequencing and the increasing amount of high throughput sequencing data [34]. Once reads are aligned, the next step is to identify germline and somatic variants in the sequenced individuals [17]. Somatic variants are novel mutations that occur within a cell population and are not inherited. Identifying these somatic variants is the first step in the process of studying mutated genes or pathways significant in a particular cancer [75]. Somatic variant hotspots can be used as therapeutic targets, as predictive cancer markers, and as indicators for cancer progression. The development of tools for identifying somatic variants has seen a rise in the past few years. Some algorithms find somatic variants by the subtraction method [76], i.e., subtracting the variants found in the normal sample from the variants identified in the tumor sample. Other tools compare a tumor sample with a matched normal sample to identify somatic variants [48]. These

tools use various parameters such as read depth, strand bias, quality score of the reads, and number of variants identified in a specific window to determine the variants.

Accuracy of identified variants can be heavily influenced by the pipelines. Depending on the purpose, for example, whether the study is of a research or clinical nature, pipelines with different strengths might be preferred. Research studies are usually conducted on multiple samples, and, therefore, false positives in individual samples will not affect overall accuracy, as presence in multiple samples is required for a variant to be identified as a true somatic variant [77]. Also, a researcher may desire the least number of false negatives to analyze all possible variants and hence may prefer a pipeline with higher sensitivity. In contrast, a clinician generally has only one sample to infer whether the sequenced individual has cancer or not and, therefore, may prefer a conservative method that has fewer false positives and thus higher precision.

With the increasing number of somatic variant callers, it is necessary to set metrics and compare the existing pipelines to recommend better parameters to guide the somatic studies. Questions about the best somatic variant caller have been popular on forums for NGS such as biostars.com [78] and seqanswers [79]. There have been studies for comparing somatic variant callers [80] but none for comparing the complete pipeline encompassing both mappers and somatic variant callers. It has been shown that read mapping affects the downstream process of germline variant calling [81]. This study evaluates the performance of all 12 combinations of three popular mappers (BWA, BWA-MEM, and Bowtie2) and four somatic variant callers (MuTect, SomaticSniper, VarScan2, and VCMM) using both real and simulated genome and

exome data. BWA [82], BWA-MEM [83], and Bowtie2 [36] use the Burrows Wheeler Transformation (BWT), which effectively uses an FM-Index to collapse the copies of a substring to align a read to the copy. The BWT-based algorithms are efficient because the reads are not aligned to just one copy of a substring and not to each copy of the substring. MuTect [84] predicts somatic variants using two Bayesian classifiers, taking into account that the variants are true mutations from the reference sequence and also are present in extremely low allele fractions in normal samples. SomaticSniper [50] uses a Bayesian comparison of genotype likelihoods between normal and tumor samples to identify somatic variants. VarScan2 [48] uses Fisher's test to differentiate germline variants from somatic variants and variants that lose heterozygosity. VCMM [85] uses a simple multinomial model to compare the probability of the variant being real variant with the probability of it being a sequencing error. To our knowledge, the study is the most comprehensive comparison of the entire pipelines for the identification of somatic point mutations to date.

2.3 Methods

2.3.1 Generating Somatic Data sets

Two data sets were generated to measure sensitivity and specificity for the somatic data sets [84]. NA12878 data set was chosen as a baseline to generate a specificity data set as it has been sequenced multiple times by different institutes using different platforms at different coverages and its variants have been catalogued and continuously updated by the National Institute of Standards and Technology (NIST) Genome in a Bottle (GIB) Consortium [81]. The specificity data set was created using two NA12878 exome data sets sequenced on Illumina

Hiseq from two different libraries (Figure 2.9). These exome data sets have a coverage of ~150X. Each data set was sorted and then divided using a custom Perl program into different bins of 5X coverage, resulting in 30 files. Since there are two NA12878 data sets, this makes a total of 60 files. Out of these 60 files, 30 files were randomly selected to form a virtual tumor specificity data set, which was used to compare with one of the original NA12878 data sets. Thus one of the original NA12878 data sets forms the normal sample while the newly formed mixture of NA12878 data sets forms the virtual tumor sample. Somatic mutations identified in this virtual tumor data set would either be germline variants that are undercalled in the normal sample or sequencing errors overcalled in the tumor sample.

The genome specificity data set was created using two NA12878 data sets sequenced on Illumina Hiseq on two different libraries with a coverage of ~30X. Each NA12878 genome data set was divided into bins of 5X coverage, partitioning the data sets into a total of 12 files. Six files were randomly selected from the 12 files to form the virtual tumor specificity genome data set. One of the original data sets was used as the normal file. The virtual tumor BAM files were converted to *fastq* using Openge. These *fastq* files were then used for running the pipelines of mapper and somatic variant calling algorithms.

To measure sensitivity, regions with heterozygous variants were identified on the NA12891 data set using three different mappers, namely, Bowtie2, BWA, and BWA-MEM with the variant caller UnifiedGenotyper of GATK. These pipelines were chosen because they had the maximum sensitivity as observed on GCAT [86]. The variants that were common to all four

variant calling pipelines were selected as heterozygous sites. We checked whether the sites were homozygous in NA12878 by comparing it against the highly confident SNP and indel call set identified by the NIST Genome in a Bottle consortium [81]. From all the reads spanning a particular homozygous region in NA12878, a randomly selected ‘x’ number of reads are replaced with heterozygous reads from NA12891 with a pre-determined allele fraction, 'af', such that af <= x (Figure 2.9). A Perl program was written to generate the sensitivity data set by replacing NA12878 reads by NA12891 reads as explained. The details of the NA12878 and NA12891 data sets used to form the sensitivity and specificity data sets for exomes and genomes are given in Supplementary Table S2.3.

2.3.2 Algorithms Used

The mappers that were used for this study include BWA [82], BWA-MEM [83] and Bowtie2 [36]. MuTect [84], SomaticSniper [50], VarScan2 [48] and Variant Caller with Multinomial probabilistic Model (VCMM) [85] were the tools used to identify somatic point mutations. The most popular mappers used by NGS users were chosen for further study based on a discussion in SEQanswers.com [79]. The most popular somatic variant callers have been used to compare in this study [80]. The version numbers are mentioned in the supplementary material (Supplementary Table S2.1). The real data sets used to simulate data in this study are mentioned in Supplementary Table S2.2.

2.3.3 Real Data

Breast invasive carcinoma exome and genome (Normal and Tumor) BAM files with IDs in Supplementary Table S2.3 were downloaded from The Cancer Genome Atlas (TCGA). These BAM files were sequenced on the Illumina platform and mapped using BWA. The variant callers MuTect, SomaticSniper, VarScan2, and VCMM were run on these 12 BAM files. The common variants should be limited to regions covered by the exome capture system. Hence, for all the genome variants, the depth in the corresponding exome sample for normal and tumor BAM files was checked. The variants that were covered by at least one base in normal and tumor BAM file were considered within regions covered by the exome capture system.

2.4 Results

Four simulated somatic data sets of exomes and genomes were generated separately to measure sensitivity and specificity. Since real data is used to simulate the data sets (see Methods section), the simulated data sets retain the base quality and the sequencing error profile of the real data. The advantage of simulating somatic data sets is that the mutation locations and expected allele fractions are already known and hence, the accuracy of the variant callers can be assessed correctly. Sites with artificially simulated somatic mutations will be called “known somatic mutations” henceforth.

2.4.1 Sensitivity and precision results for pipelines that detect somatic point mutations for exomes

The known somatic mutations predicted by a somatic variant caller are true positives (TP), and the ones not predicted by the somatic variant caller are false negatives (FN).

Sensitivity was calculated using the formula TP/(TP+FN). Figure 2.1 shows the sensitivity of different pipelines for detecting somatic point mutations. In general, irrespective of the mapper used, the somatic variant caller VCMM has the highest sensitivity. VCMM, MuTect, VarScan2, and SomaticSniper have decreasing sensitivity in that order. The average standard deviation of a pipeline's sensitivity within somatic variant callers decreases by 96% when compared to standard deviation within mappers, and the average sensitivity of pipelines within mappers is 0.19, 0.16, and 0.15 for BWA, BWA-MEM, and Bowtie2, respectively. Thus, the mapper does not affect the sensitivity as much as the somatic variant caller (Figure 2.1).

Precision was calculated using the formula TP/(TP+FP). The TPs were calculated from the sensitivity data set while the FPs were considered from the specificity data set. MuTect has the highest average precision (0.68), followed by VarScan2 (0.27), VCMM (0.02), and SomaticSniper (0.009) (Figure 2.1).

We also calculated the F1 score that combines sensitivity and precision to give the overall metric for evaluating the performance of the methods. This score is calculated using the formula (2*precision*recall)/(precision+recall). For exome data, MuTect has the highest F1 score followed by VarScan2, VCMM, and SomaticSniper (Figure 2.1).

In the case of the specificity data set, any detected variant is a false positive (FP), since the specificity data set is a mixture of two data sets of the same individual. This variant detection could be a result of germline variants being undercalled in the normal sample or sequencing

errors being overcalled in the tumor sample. All the sites that are not called as variants are true negatives (TN) in this case. Specificity was calculated using the formula TN/(TN+FP). All the pipelines have a high specificity, ranging from 99.9986 to 99.9989% (results not shown for brevity). The specificity of all the methods is high because the number of TNs is extremely high, much higher than the number of FPs and as a result, dominates the formula TN/(TN+FP). MuTect has the highest specificity, although the difference in specificity from other calls is very small (~0.0001). The significance of the difference in FPs is studied further.

Supplementary Figure S2.1 shows the number of FPs and FNs detected by different methods in exomes. VCMM detects the highest number of FPs. The number of FPs decreases from SomaticSniper to VarScan2, and MuTect is the most conservative of all the somatic variant callers. The number of FPs range from 354 to 648 using MuTect, while the number ranges from 95,100 to 113,527 using VCMM. The average reduction in FPs from VCMM to SomaticSniper, VarScan2, and MuTect is 21.5%, 96.3%, and 99.6% respectively, indicating that MuTect and VarScan2 are quite conservative in their approaches. The number of FNs identified decreases from SomaticSniper to VarScan2, MuTect, and VCMM in that order. The average reduction in the number of FNs from SomaticSniper to MuTect and VarScan2 is 1.4% and 0.7%, respectively. The average reduction in FNs from SomaticSniper to VCMM is 23.4%. This indicates that there is a significant decrease in the number of FNs in VCMM, indicating that the sensitivity is high in VCMM.

2.4.2 Sensitivity and precision results for pipelines that detect somatic point mutations for genomes

Since genomes are usually sequenced at depths of 30X - 60X, the performance of the pipelines on genomes might differ from that on exomes that are usually covered at much higher sequencing depths of 100X-200X. Hence, the virtual tumor sensitivity and specificity data sets were created for genomes as well. This is also a good exercise to find if different data sets affect the performance of pipelines. Figure 2.2 shows the sensitivity of different pipelines in detecting somatic point mutations for genomes. The standard deviation of a pipeline within a mapper is 0.29, 0.30, and 0.29 for Bowtie, BWA-MEM, and BWA respectively. The standard deviation within a somatic variant caller is 0.001, 0.01, 0.003, and 0.007 for MuTect, VarScan2, SomaticSniper, and VCMM respectively. There is a 75% decrease in the average standard variation within mappers to average standard variation within somatic variant callers in genomes, lesser than the standard deviation decrease in exomes. This indicates that identification of somatic variants is affected more by mappers for genome data than for exome data. For sensitivity, the trend of the pipelines for genomes is similar to that seen in exome data, i.e., VCMM still has the highest sensitivity and MuTect is the most conservative somatic point mutation caller. VarScan2 and SomaticSniper are somatic callers that identify an intermediate number of somatic point mutations with VarScan2 detecting fewer than SomaticSniper. For precision, SomaticSniper has the highest precision, with an average precision 0.0006 (Figure 2.2), followed by MuTect, VarScan2, and VCMM. The F1 score is the highest for SomaticSniper, followed by MuTect, VCMM, and VarScan2 (Figure 2.2).

All pipelines have a high specificity ranging from 0.9985 (VCMM) to 0.9999 (MuTect) (results not shown). VCMM detects the highest number of FPs (Supplementary Figure S2.2). The number of FPs generally decreases from SomaticSniper to VarScan2 with an average drop of 45.84% and MuTect predicts the least number of FPs at an average of 16,628. The FPs predicted by MuTect drops by 99% from those predicted by VCMM, suggesting that MuTect is extremely conservative in its approach. The number of FNs identified decreases from MuTect to VarScan2, SomaticSniper, and VCMM, in that order. VarScan2 and SomaticSniper predict an intermediate number of FPs and FNs with MuTect being very conservative and VCMM being the most liberal (Supplementary Figure 2.2). FNs predicted by VCMM drops by 62% from those predicted by MuTect.

2.4.3 Causes for undetected true somatic variants by pipelines in exomes

As VCMM has the highest sensitivity among the variant callers, we examined it further to see whether changing its parameters could reduce the number of false positives and false negatives. We found that irrespective of the change in different parameters, VCMM failed to detect certain somatic variants (results not shown here).

The causes for the failure to detect somatic variants by different algorithms were further examined. The analysis revealed that one of the reasons variants were not detected was due to regions being heterozygous for the normal sample (NA12878). The sensitivity data set was created by mixing reads that were homozygous for NA12878 and heterozygous for NA12891. The list of homozygous regions was derived from the catalog of variant sites provided by NIST.

Since the catalog provided by NIST was a result of the analysis of 14 data sets, the regions that are heterozygous in the NA12878 sample that we used could be a result of sequencing error or a result of variants not yet updated by NIST. So, the FNs that are detected by different pipelines could have been identified correctly if the normal sample was sequenced correctly. Other factors contributing to variants not being detected by pipelines include (i) the strand bias was less than 10%, (ii) the sum of all the alternate alleles' quality scores is less than 60, (iii) the normal sample did not have adequate coverage, and (iv) the somatic variant did not have adequate coverage in the tumor sample. These factors are the result of limitations in sequencing, mapping, or somatic variant calling. The reason that the number 10% was chosen for strand bias and 60 was chosen for alternate alleles' quality score was because most somatic variant calling algorithms use these cutoff values as their parameters [48] . Figure 2.3 shows a distribution of the factors due to which certain somatic sites are undetected. The undetected somatic variants by a somatic variant caller failing to fall under any of the above categories are mentioned under the category called others. Other factors could include low mapping quality, and low base quality for bases covering the variant position.

All the somatic sites that fall under the aforementioned factors were removed and then the sensitivity was checked again (Figure 2.4). The reason for this is we wanted to check the performance of the pipelines in an ideal case scenario where the bases were sequenced at a high quality, with a high coverage and accurately. This set of somatic variants is called a high quality data set. The sensitivity of the pipelines identifying somatic point mutations increases by at least 51% in this scenario. The average sensitivity increase for MuTect, VarScan2, SomaticSniper, and

VCMM are 83.5%, 76.72%, 77.68%, and 51.42%, respectively. The best pipeline BWA+VCMM correctly predicts 94.25% of the somatic point mutations in this case.

2.4.5 Sensitivity and precision vs. tumor sequencing depth for exomes and genomes

Figure 2.5 shows the sensitivity and precision of different pipelines as a function of tumor sample sequencing depth for exome data. For all pipelines, both sensitivity and precision increase with the increase of sequencing depth, although the increase is much greater initially when sequencing depth goes from 10x to 20x. For all sequencing depth, VCMM has the highest sensitivity, followed by other callers with similar sensitivity. For precision, MuTect performs the best, followed by VarScan2, VCMM, and SomaticSniper. The pattern for the genome data is similar to the observation here for the exome data (results not shown for brevity).

2.4.6 Sensitivity vs. allele fraction for exomes and genomes

Sensitivity of different pipelines as a function of allele fraction for exomes is displayed in Figure 2.6. VCMM performs the best irrespective of the allelic fraction for both exome data and genome data. MuTect performs the second best for exomes but the worst for genomes. At lower allele fractions the performance of VarScan2 decreases by 62% when compared to MuTect. SomaticSniper and VarScan2 perform similarly for both exomes and genomes.

2.4.7 Comparing pipelines using real data

To assess the best variant caller using real data, the four variant callers were run on three breast cancer normal-tumor samples of exomes and genomes (details in Methods). It is expected that the best variant caller would predict the highest proportion of shared variants between the exome and the genome sequenced from the same individual. To make a fair comparison, only variants that were covered by at least one read in both normal and tumor samples for both whole genome and whole exome platforms were considered. Thus, we made sure that only variants from exonic regions of the genomes were compared with the variants from the exomes. Figure 2.7 shows the mean percentage concordance of somatic variants between whole exome data (WXS) and whole genome data (WGS). VCMM identifies the highest percentage (11%) of common variants, followed by MuTect (8%), SomaticSniper (4%), and VarScan2 (2%). However, the small proportion of called variants shared by exome data and genome data shows that there is much room for improvement in somatic variant calling.

2.4.8 Number of germline variants misidentified as somatic variants

Germline variants that were falsely identified as somatic variants in real data were analyzed. GATK HaplotypeCaller was used to identify germline variants in the normal samples because it has been shown to have high precision in germline variant calling when compared to other variant callers, GATK UNifiedGenotyper, Samtools, Atlas, and Freebayes [87]. Figure 2.8 shows the average percentage of germline variants predicted by different variant callers. The percentage of germline variants misidentified as somatic variants by different pipelines were similar in case of exomes and genomes. Approximately 0.08%, 4.5%, and 6% of the variants

called by MuTect, SomaticSniper, and VarScan2 were actually germline variants, as compared to 77-85% called by VCMM. The high percentages called by VCMM is mainly due to the fact that VCMM does not consider whether a variant is germline or not during calling whereas the other three tools do.

2.5 Discussion

In general, standard deviation within variant callers is higher than standard deviation within mappers for both exome data and genome data, suggesting that callers affect variant calling more than mapping. The reason for this could be that approaches used by different somatic variant callers can vary a great deal and many different filter parameters are also deployed in the process to finally call a variant. For example, VCMM, MuTect, VarScan2, and SomaticSniper all have different cutoff values for parameters such as minimum depth in normal and tumor samples, minimum allele fraction in tumor sample, strand bias, minimum base quality, and mapping quality. MuTect [84] penalizes heavily the presence of a variant in the normal sample, VarScan2 [48] and SomaticSniper [50] were developed to be more lenient towards the presence of variants in normal samples whereas VCMM [85] was developed to identify somatic variants from only tumor samples. All of these aspects affect the results of somatic variant calling. Compared to variant callers, mappers use the same seed-and-extend strategy, and identification of the seed for extension is done either by hashing or Burrows Wheeler transform (BWT). In this case all the selected popular mappers (BWA, BWA-MEM, Bowtie2) use BWT. Moreover, there are fewer parameters and filters considered for mapping than for calling.

Comparison of exome data and genome data suggests that mappers affect the identification of somatic variants more in genomes than in exomes (Figure 2.1 vs. Figure 2.2). This makes intuitive sense as usually exome data tend to have much higher coverage (~100X) than genome data, which has a low coverage (~30X) due to the cost associated with higher coverage for an entire genome. Mapped regions will not be affected by slight variations in exomes due to the higher coverage, whereas for genomes mapped regions will likely be more affected by the mapper due to the low coverage, which in turn influences more downstream variant calling.

Sequencing depth has been known to be an important factor affecting the performance of both germline [87] and somatic variant callers [88] and as expected, callers usually perform better at higher sequencing depths than low ones. Similarly, sequencing depth has a positive impact on mapping quality, that is, the higher the sequencing depth, the better the mapping quality. The current comparison of the twelve combinations of mappers and callers show a similar pattern (Figure 2.5). For sequencing depths from 10x to 80x for both exomes and genomes, the trend is similar among the pipelines (Figure 2.1 and Figure 2.2). VCMM has the highest sensitivity followed by SomaticSniper, VarScan2, and MuTect. MuTect has the highest precision, followed by VarScan2, VCMM, and SomaticSniper. MuTect consistently has high precision values in the case of exomes and genomes while VCMM does not perform well in terms of precision. It is interesting to note that SomaticSniper, which has the lowest precision for exome data, has the highest precision values in the case of genomes. This could be due to the cross-contamination of normal and tumor samples being considered. At lower allele fractions,

other than VCMM, MuTect has the highest sensitivity. We expect VCMM to do well since it falsely identifies a lot of germline variants as somatic variants. VCMM performs the best irrespective of the allele fraction. MuTect too performs well at lower allele fractions as has been seen by previous studies [84]. All the methods perform well at higher allele fractions similar to the results in previous studies [84].

When a high quality data set is considered, there is a spike in performance for each pipeline. The performance of each pipeline increases by at least 51% when the high quality data set is considered. The highest spike in performance is by MuTect, although the best performing variant caller in terms of sensitivity is still VCMM. This spike in performance is due to the ideal case scenario being considered wherein there is no strand bias, the total depth is higher than 12, the allele fraction is greater than 10%, the sum of allele base quality is greater than 60, and the base quality for each base is greater than 30. These cutoffs are frequently used by most somatic variant callers for their parameters. Our sensitivity and precision results are much lower than the results found in previous work comparing somatic variant callers [88], which could be due to the fact that we simulated the data set using real data sets and took into account different allele fractions and cross-contamination in our simulated data sets. But once we consider a high quality data set, our results are similar to the results seen previously [88].

For both exomes and genomes, VCMM predicts the greatest number of somatic variants and at least 20 times the variants that the other somatic variant callers predict. VarScan2 identified more somatic variants than MuTect which is similar to a previous study [88].

Approximately 80% of the variants that VCMM predicts are germline variants. The reason for this could be that VCMM, unlike VarScan2, MuTect, and SomaticSniper, does not take into account the normal samples while identifying somatic variants. In contrast, MuTect identifies fewer than 1% of germline variants as somatic variants. This can be attributed to the fact that MuTect heavily penalizes the presence of an allele in normal samples [84].

An analysis of the concordance of called somatic variants in exomes and genomes within the exonic regions over three samples of data from TCGA was done. The analysis revealed that the concordance is only 11% at best, suggesting that there is much scope for improvement for somatic variant callers. Similar to these current results, low concordance was also observed for called germline variants in the exome data and the exonic regions of the whole genome data [89].

So essentially, a user can decide between VCMM and MuTect according to their specific requirement. Our analysis of genomes also corroborates our findings from the analysis of somatic exome data that using any of the popular mappers does not affect the identification of somatic variants. VCMM can be used as a somatic variant caller when multiple normal-tumor samples are present and MuTect is the most conservative somatic variant caller.

2.6 Conclusion

The decrease in the average standard deviation within somatic variant caller to average standard variation within mapper shows that using any of the mappers mentioned in this paper should work fine for identification of somatic variants. VCMM has the highest sensitivity among

all the different variant callers used to detect somatic point mutations. VCMM detects a high number of false positives and approximately 80% of the somatic variants it predicts are germline variants. MuTect is the most conservative method to detect somatic point mutations. At lower allele fractions, VCMM and MuTect perform the best in the case of exomes and genomes. If a user would like to maintain a good balance between the number of false positives and false negatives in exomes and genomes, using MuTect would be a safe bet.

2.7 Abbreviations

FNs: False Negatives

FPs: False Positives

GIB: Genome In a Bottle

NIST: National Institute of Standards and Technology

TCGA: The Cancer Genome Atlas

TNs: True Negatives

TPs: True Positives

WGS: Whole Genome Sequence

WXS: Whole Exome Sequence

Figures

Figure 2.1: Sensitivity, precision and F1 score of different pipelines in detecting somatic point mutations in exomes.

SS denotes SomaticSniper for all figures. V2 denotes VarScan2 for all figures.

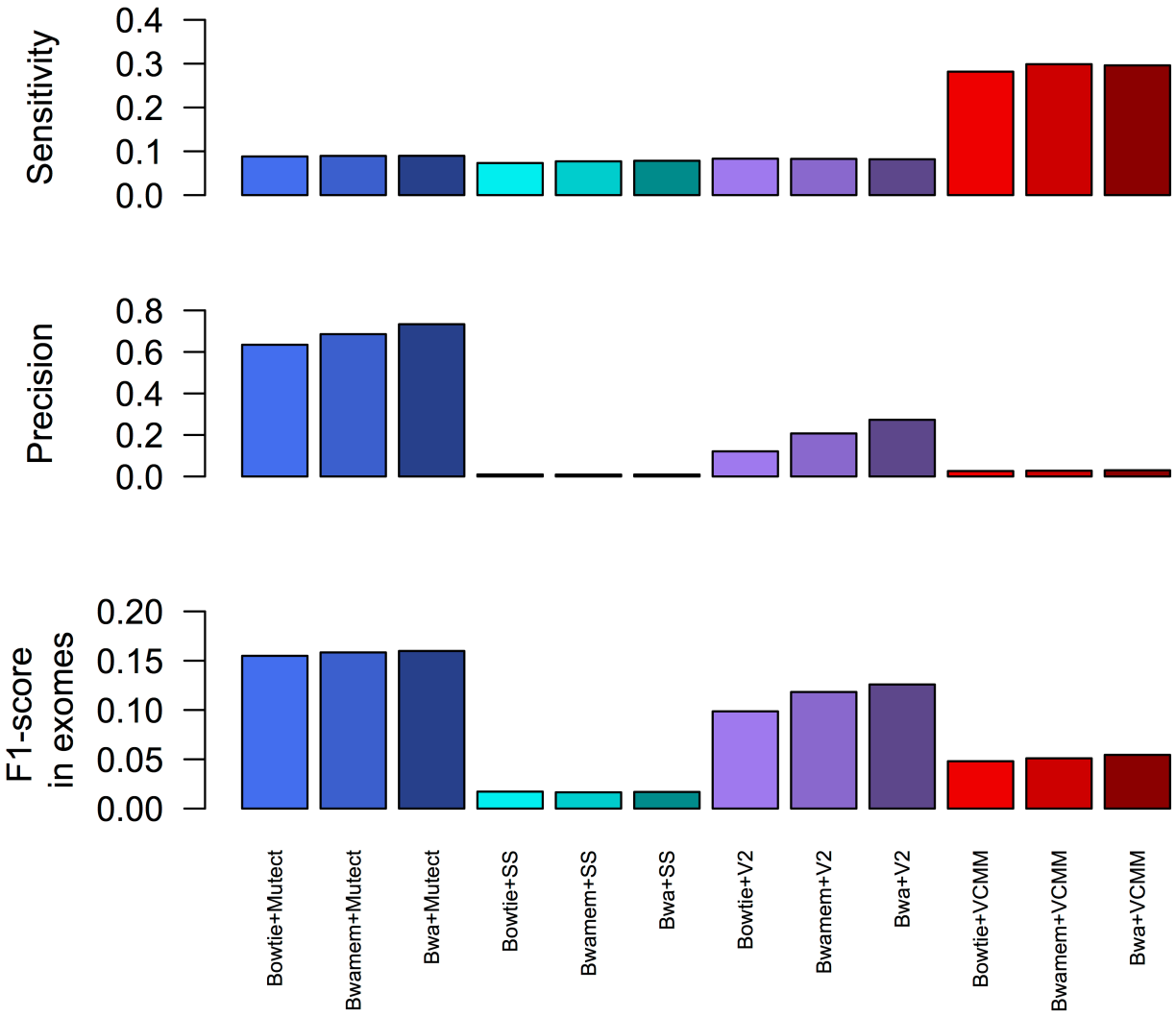


Figure 2.2: Sensitivity, precision and F1 score of different pipelines detecting somatic point mutations in genomes.

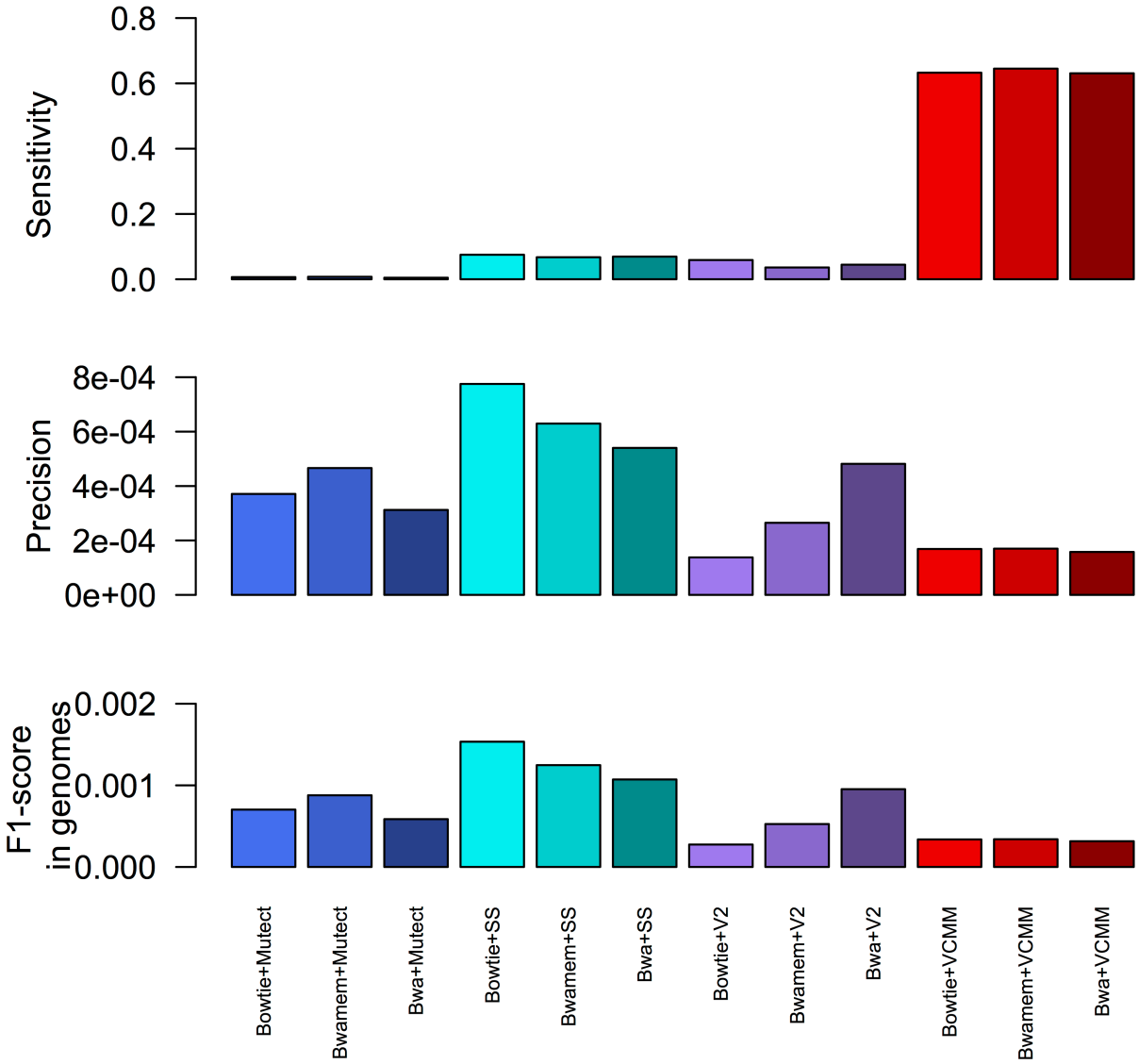


Figure 2.3: Factors affecting the detection of somatic point mutations.

Het- The site is heterozygous in the normal sample; AQS (Allele Quality Score)- Sum of all the alternate bases' quality score is lesser than 60; SB (Strand Bias)- Strand Bias is lesser than 10%; Depth- Total depth is lesser than 12 or depth of base is lesser than 3; Others- Present in low complexity regions, low mapping quality, low sequencing depth in Normal sample

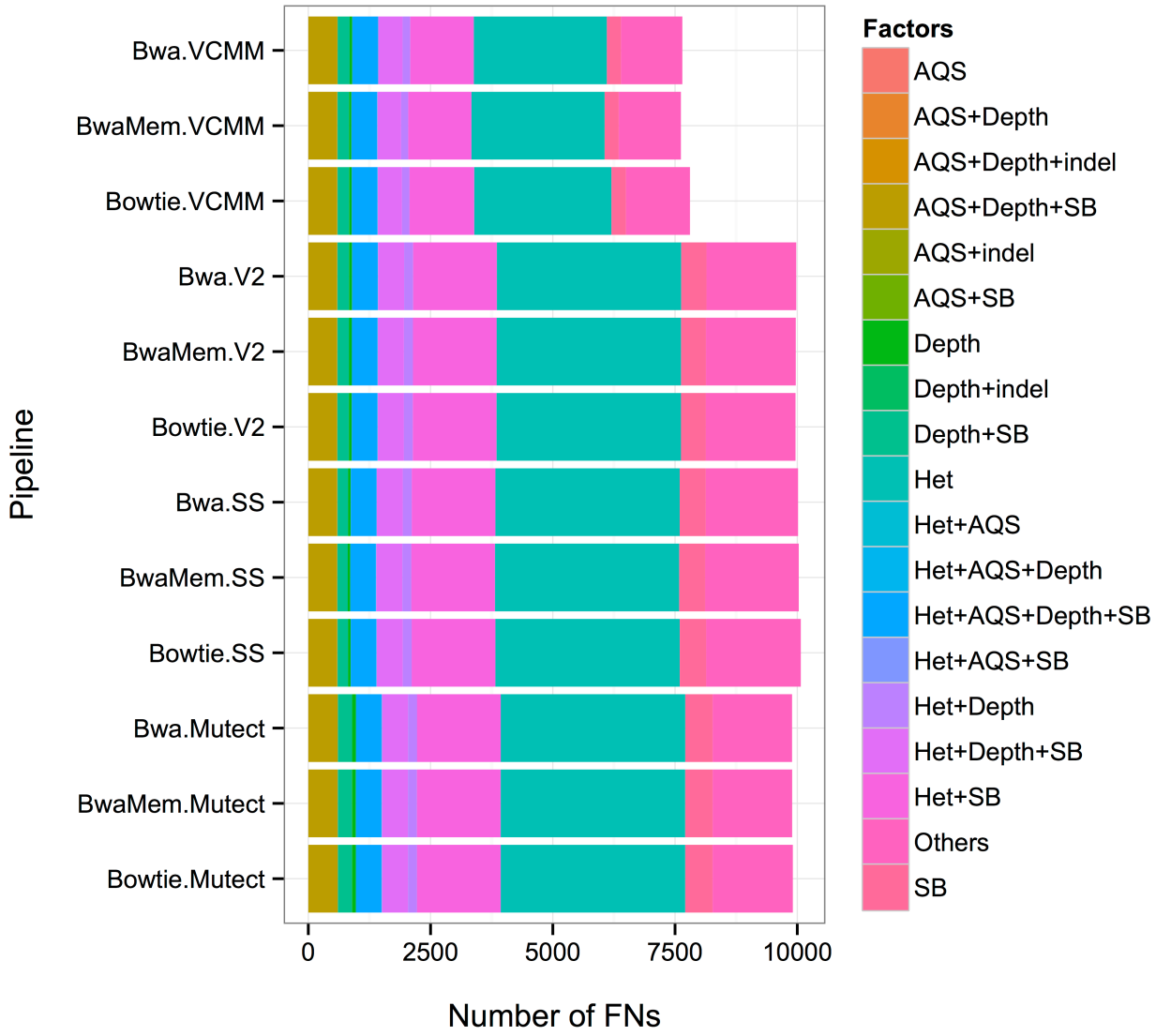


Figure 2.4: Sensitivity of different pipelines in detecting somatic point mutations using a high quality exome data set.

The high quality data set is constructed by removing (i) Somatic pint mutations that are in sites that are heterozygous in the normal sample (ii) Sum of all the alternate bases' quality score lesser than 60 (iii) Strand Bias lesser than 10% (iv) Total depth lesser than 12 or depth of allele lesser than 3.

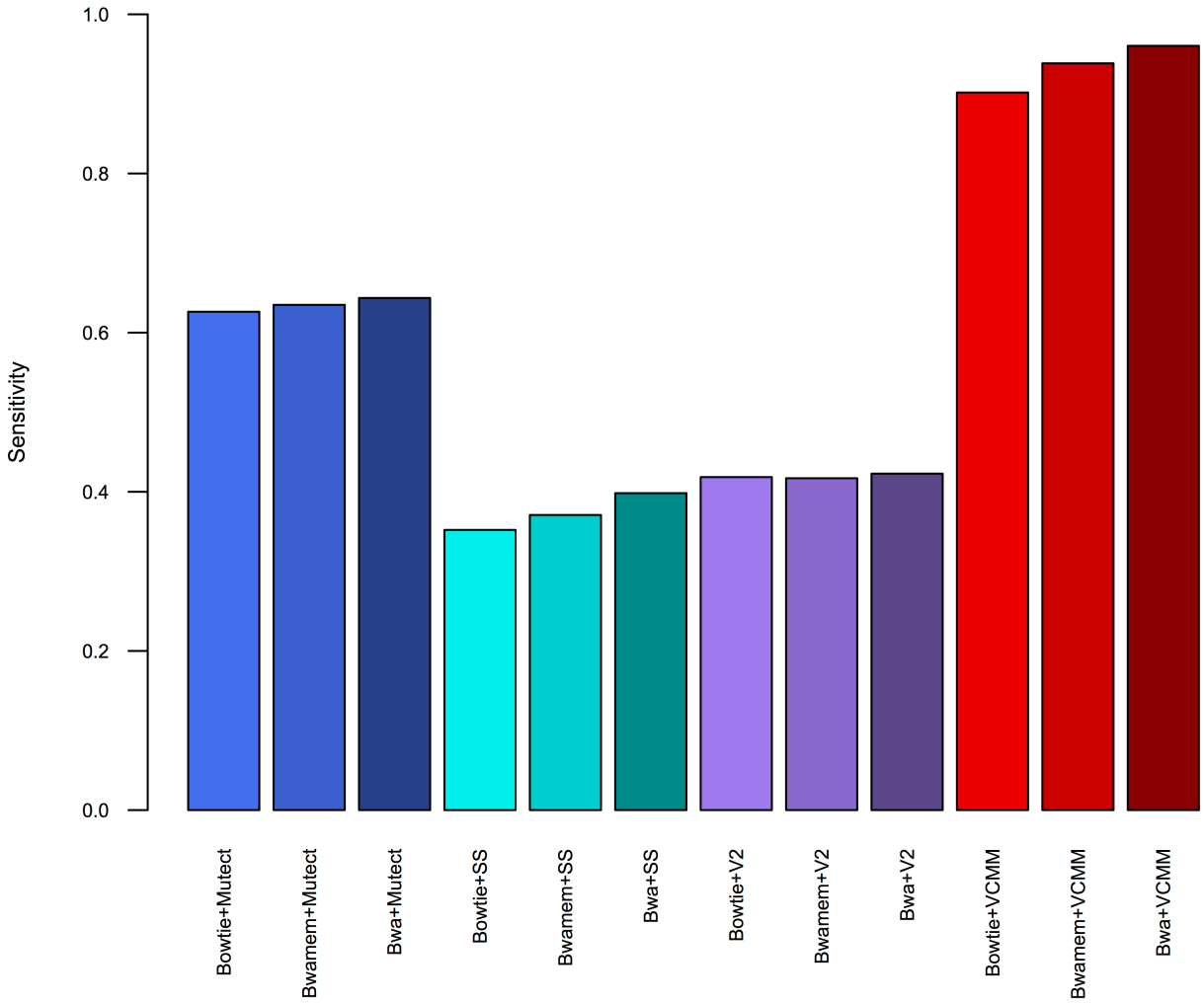


Figure 2.5: Sensitivity and precision as a function of tumor sequencing depth of different pipelines while detecting somatic point mutations in exomes.

Shades of Red indicate VCMM, shades of green indicate VarScan2, shades of blue indicate SomaticSniper an shades of purple indicate MuTect. Shades from dark to light refer to BWA, BWA-MEM and Bowtie.

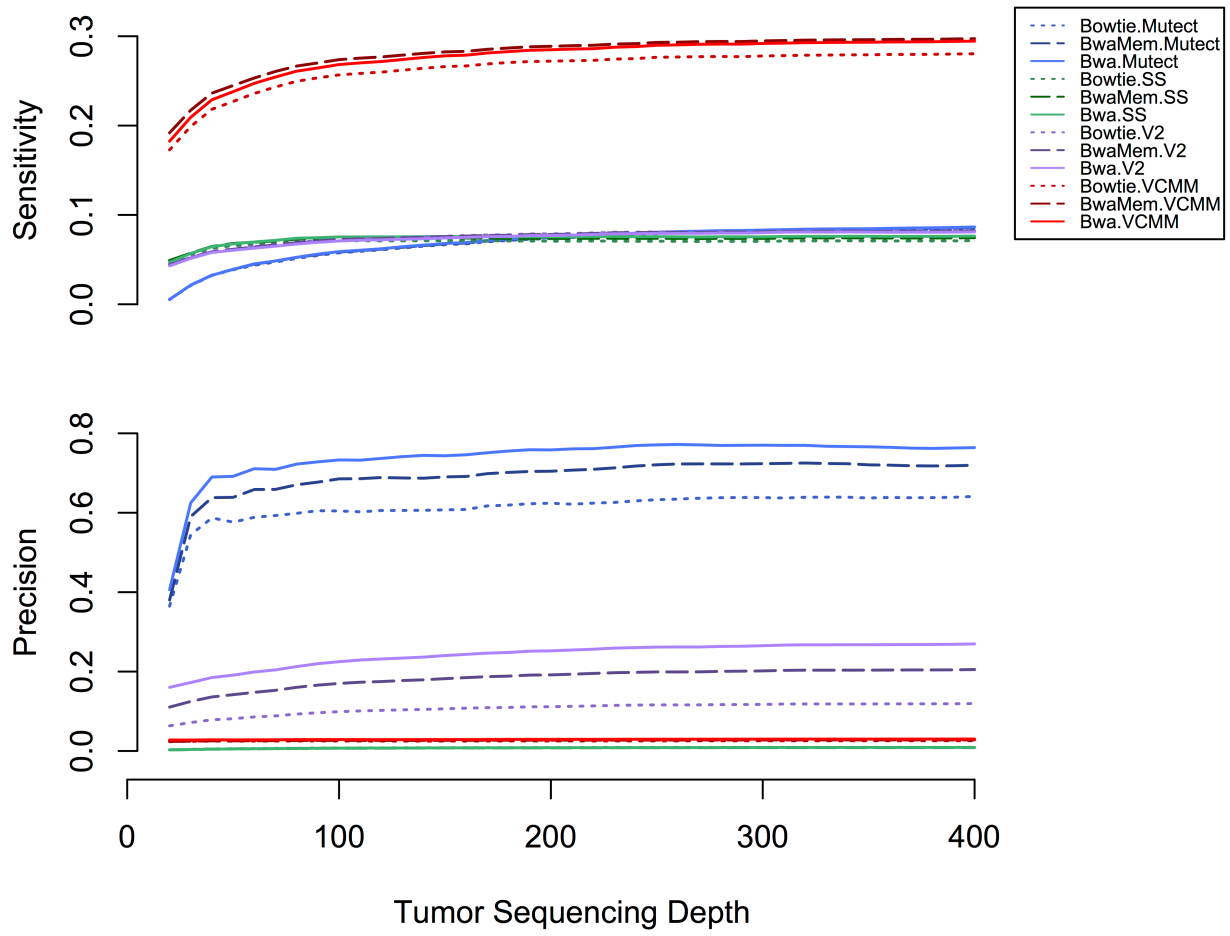


Figure 2.6: Sensitivity as a function of tumor allele fraction while detecting somatic point mutations in exomes.

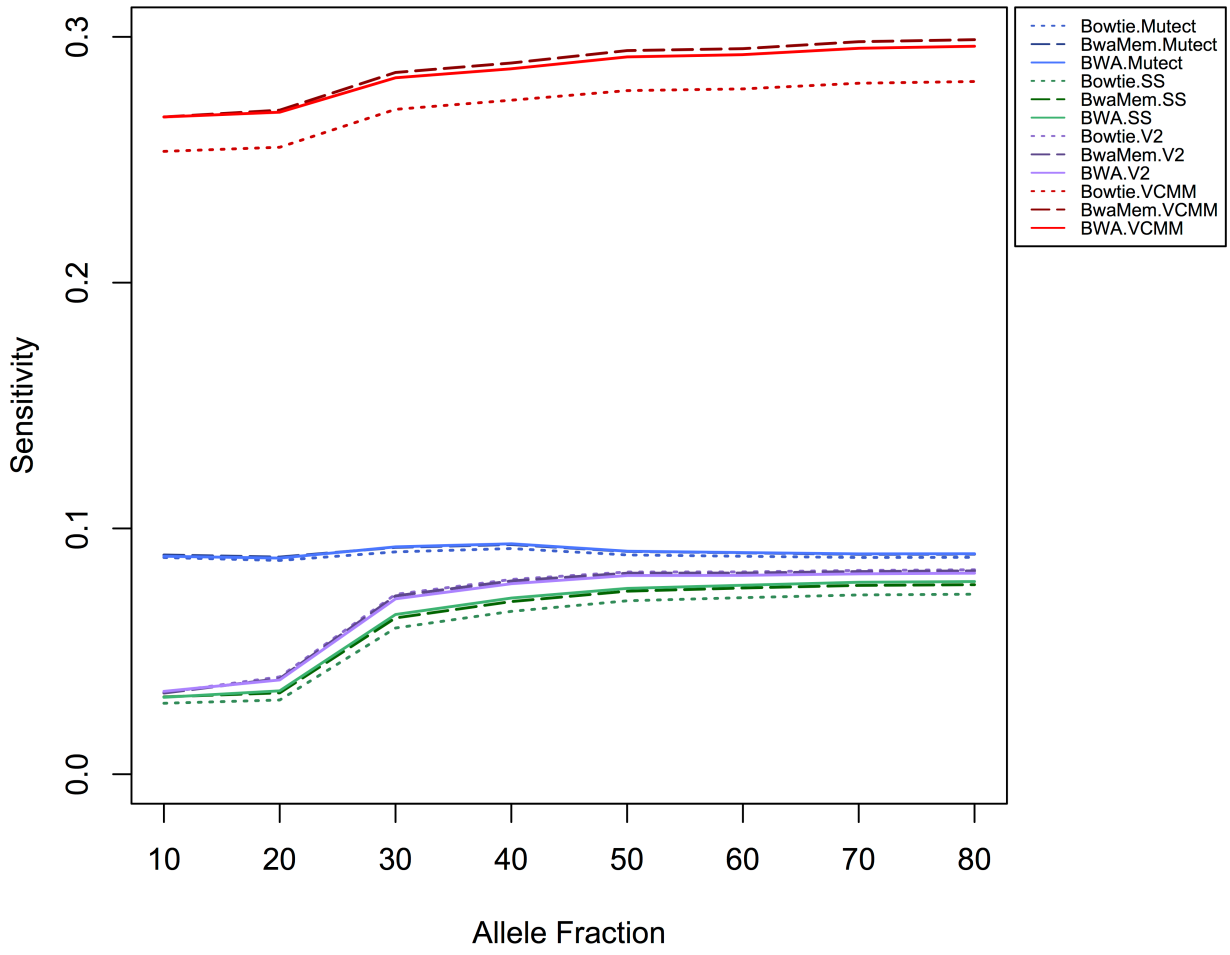


Figure 2.7: Mean concordance (in percentages) of somatic variants over three samples of real data between whole exome sequences (WXS) and whole genome sequences (WGS).

The intersection between WXS and WGS is the percentage of variants called by both types of data.

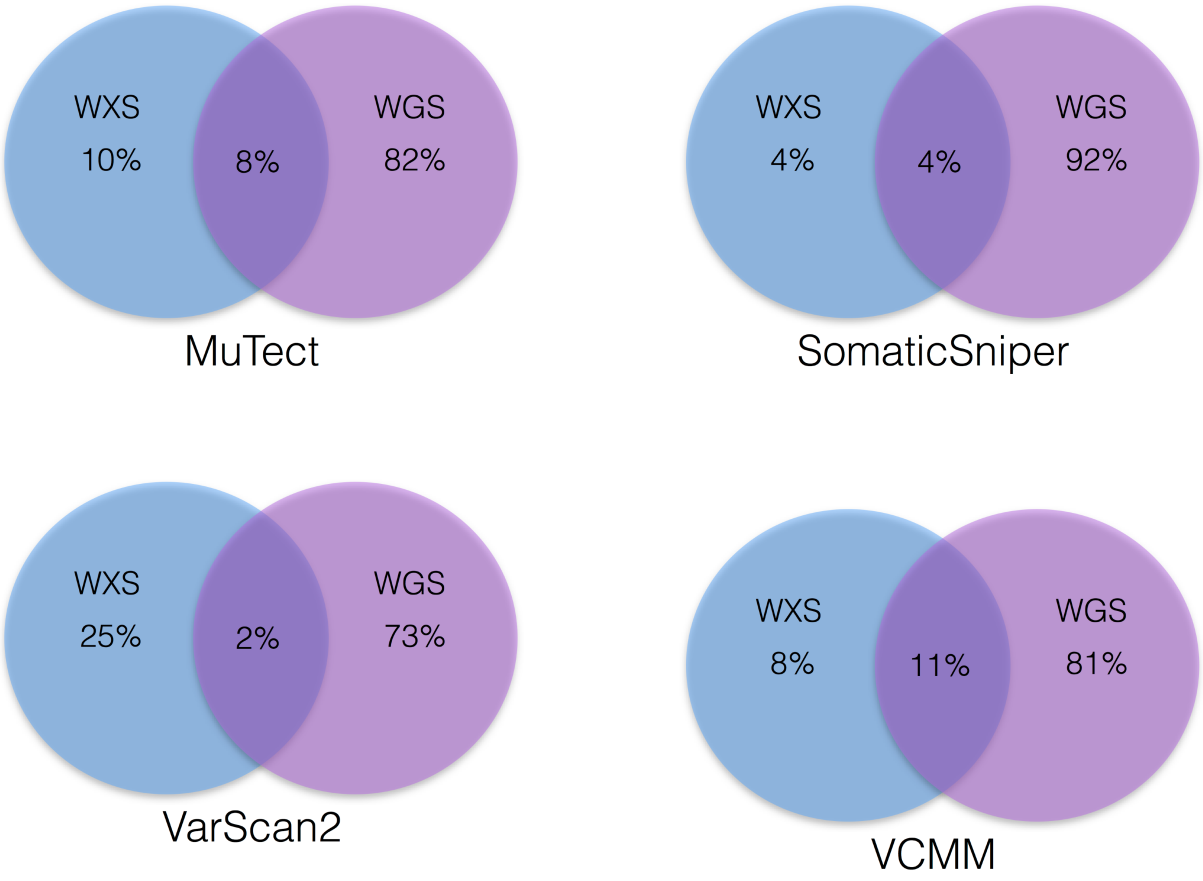


Figure 2.8: Percentage of germline variants misidentified as somatic variants in real data sets.

a) In case of exomes b) In case of genomes. The numbers indicate percentage

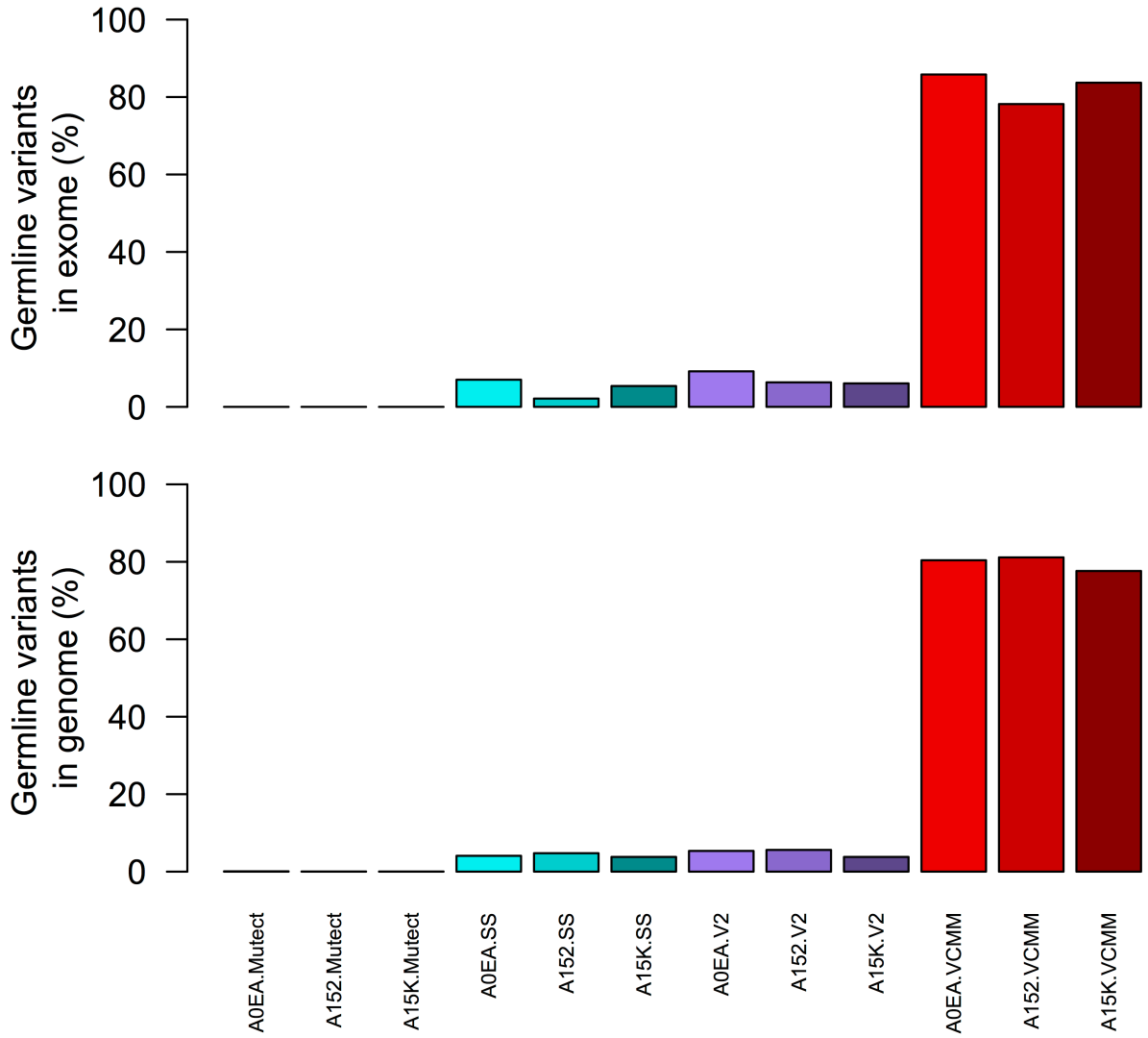
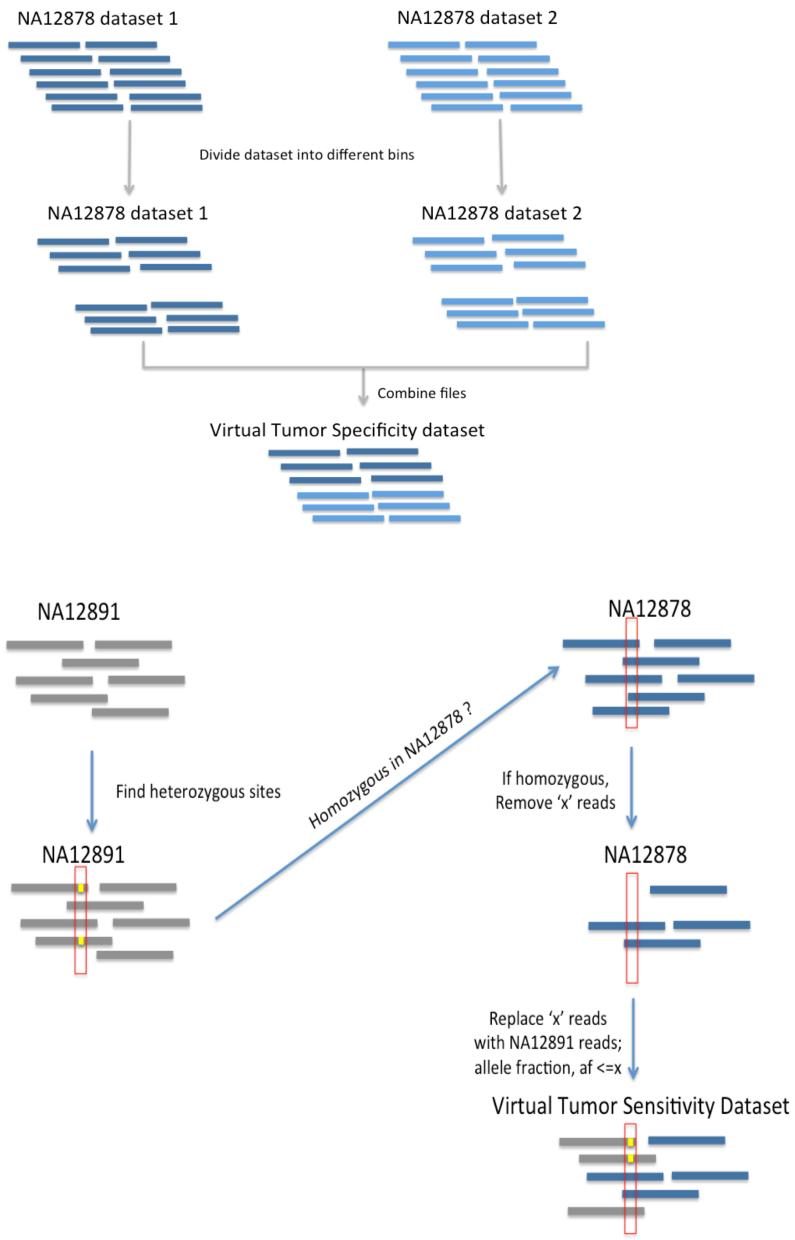


Figure 2.9 - Generating sensitivity and specificity data set



Supplementary Figures

Figure S2.1: False positives and false negatives identified by different pipelines while detecting somatic point mutations for exomes.

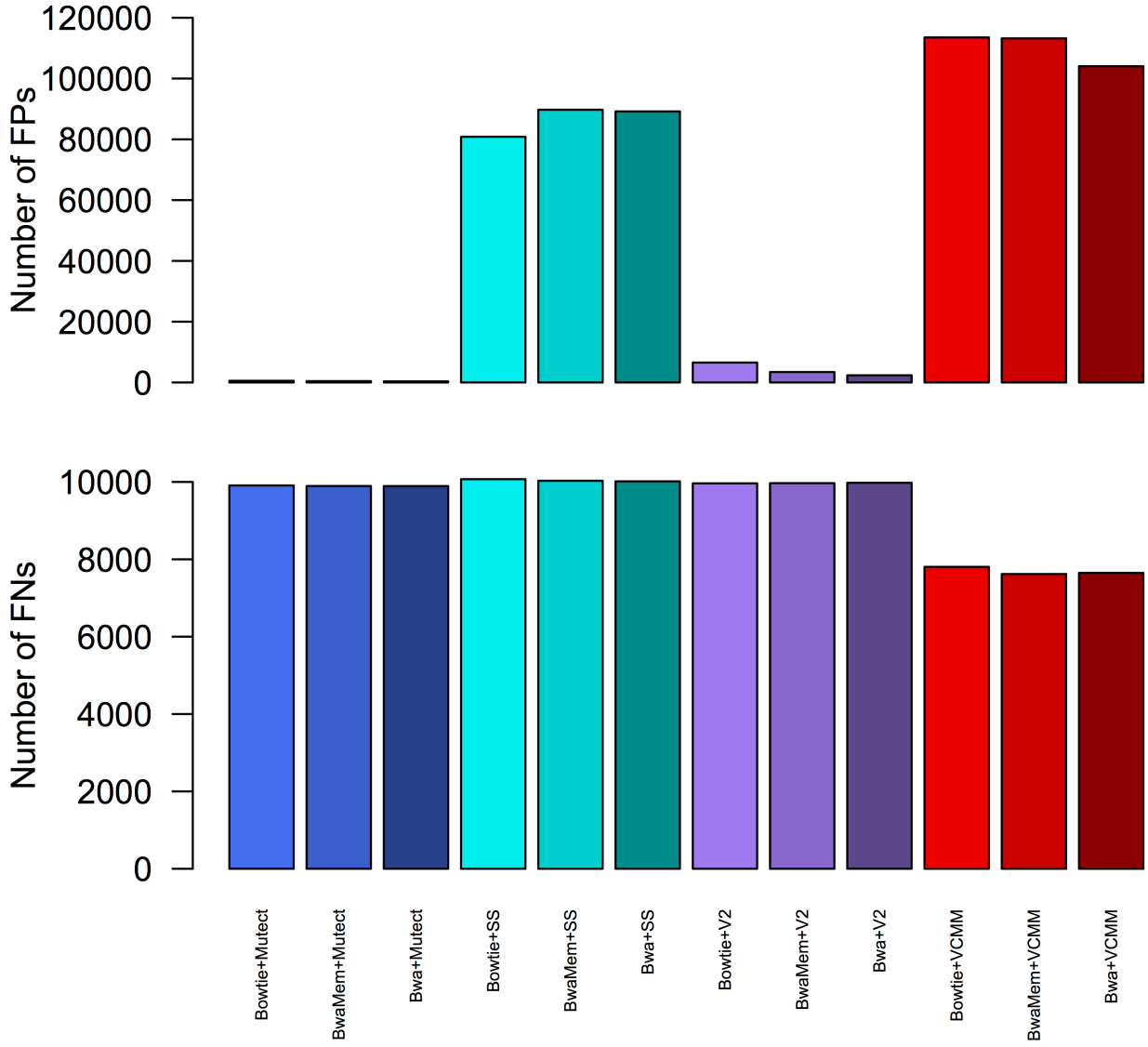
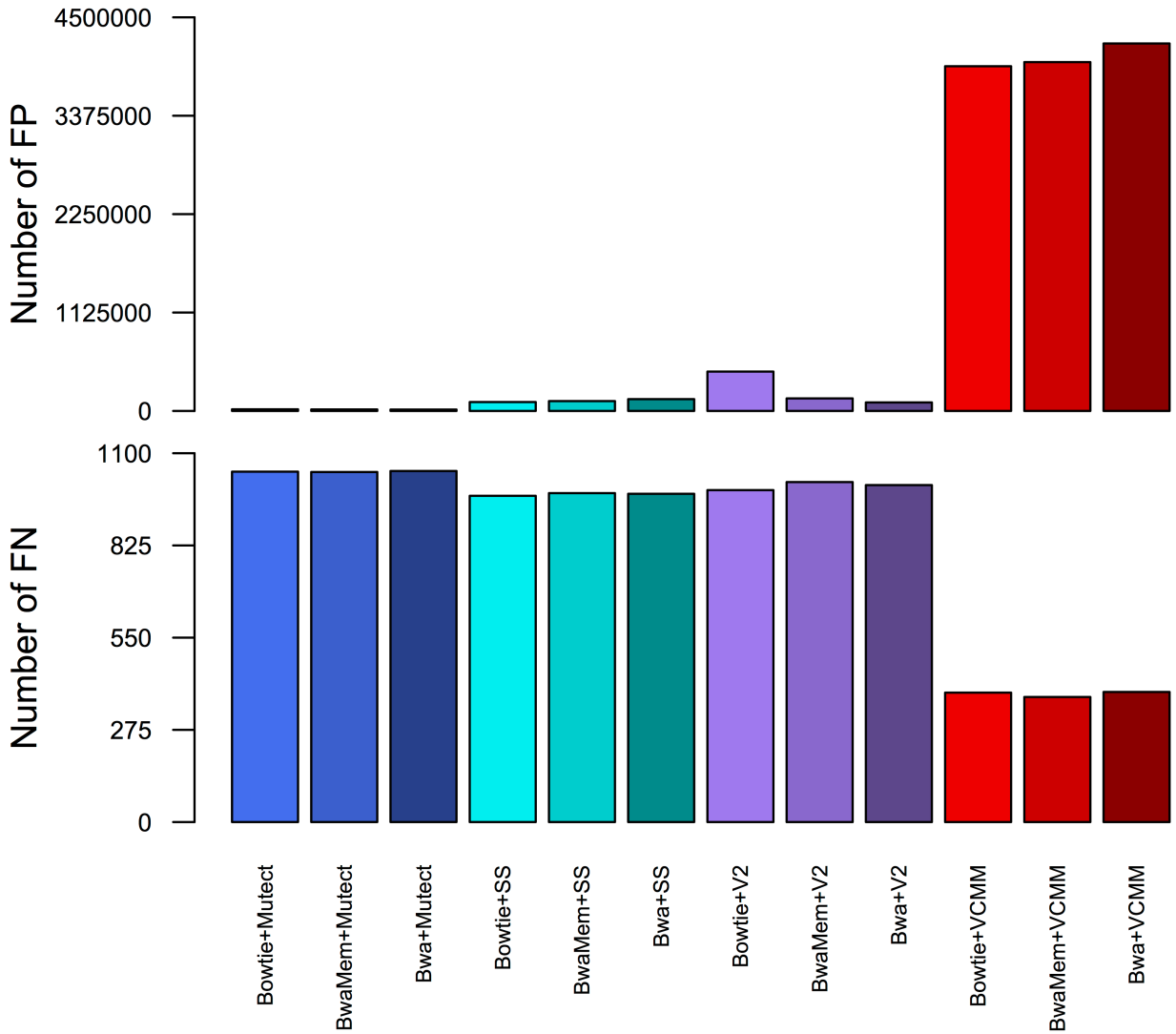


Figure S2.2: Number of false positives and false negatives identified by the pipelines while detecting somatic point mutations for genomes.



Supplementary Tables

Table S2.1: The version number of each software used in the study

Tool	Version Number
Bowtie2	2.1.0
BWA	0.7.10-r789
BWA-MEM	0.7.10-r789
MuTect	1.1.7
SomaticSniper	1.0.4
VarScan2	2.3.6
VCMM	1.0.1

Table S2.2: Details of the real data sets used for simulation

Data set	Type	Read Length	Platform	URL
NA12878	WXS	76	Illumina HiSeq	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_ceu_trio_b37_decoy/
NA12878	WXS	76	Illumina HiSeq	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_ceu_trio_b37_decoy/
NA12891	WXS	76	Illumina HiSeq	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_ceu_trio_b37_decoy/
NA12878	WGS	150	Illumina HiSeq	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_ceu_trio_b37_decoy/
NA12878	WGS	150	Illumina HiSeq	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_ceu_trio_b37_decoy/
NA12891	WGS	150	Illumina HiSeq	ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_ceu_trio_b37_decoy/

Table S2.3: Real data samples used in the study.

NB refers to Blood Derived Normal, TP Primary solid Tumor. WXS refers to whole exome data, WGS whole genome data.

Barcode	Sample Type	Library Type
TCGA-BH-A0EA-10A-01D-A110-09	NB	WXS
TCGA-BH-A0EA-01A-11D-A10Y-09	TP	WXS
TCGA-BH-A0EA-10A-01D-A110-09	NB	WGS
TCGA-BH-A0EA-01A-11D-A314-09	TP	WGS
TCGA-E2-A152-10A-01D-A12B-09	NB	WXS
TCGA-E2-A152-01A-11D-A12B-09	TP	WXS
TCGA-E2-A152-10A-01D-A19H-09	NB	WGS
TCGA-E2-A152-01A-11D-A19H-09	TP	WGS
TCGA-E2-A15K-10A-01D-A12Q-09	NB	WXS
TCGA-E2-A15K-01A-11D-A12Q-09	TP	WXS
TCGA-E2-A15K-10A-01D-A12Q-09	NB	WGS
TCGA-E2-A15K-01A-11D-A314-09	TP	WGS

Chapter Three

Framework for integration of genome and exome data to improve the identification of somatic variants

3.1 Abstract

Cost-effective high-throughput sequencing technologies, together with efficient mapping and variant calling tools, have made it possible to identify somatic variants for cancer study. However, integrating somatic variants from whole exome and whole genome studies poses a challenge to researchers, as the variants identified by whole genome analysis may not be identified by whole exome analysis and vice versa. Simply taking the union or intersection of the results may lead to too many false positives or too many false negatives. To tackle this problem, we use machine learning models to integrate whole exome and whole genome calling results from two representative tools, VCMM (with the highest sensitivity but very low precision) and MuTect (with the highest precision). The evaluation results, based on both simulated and real data, show that our framework improves somatic variant calling and is more accurate in identifying somatic variants than either individual method used alone or using variants identified from only whole genome data or only whole exome data. Using a machine learning approach to combine results from multiple calling methods on multiple data platforms (e.g., genome and exome) enables more accurate identification of somatic variants.

Keywords: Somatic variants, genome and exome analysis, framework for combining results from tools

3.2 Introduction

Somatic variants, unlike germline variants, are novel mutations that occur within a cell population and are not inherited. Identification of somatic variants enables the identification of variant hotspots. These hotspots can be used to study significant genes and pathways that can then be used in predictive, prognostic, remission and metastatic analysis of cancer. These somatic variant hotspots can also be used as therapeutic targets. Identifying somatic variants is more difficult than identifying germline variants because of copy number aberrations and the variability of somatic mutations.

In the past few years, a lot of methods have been developed to identify somatic variants. These programs differ in the kinds of statistics used and the parameters considered. For instance, SomaticSniper [90] uses a Bayesian approach to identify somatic variants. VarScan2 [91] uses Fisher's test to differentiate germline variants from somatic variants and variants that lose heterozygosity. MuTect [92] predicts somatic variants by two Bayesian classifiers, taking into account that the variants are true mutations from the reference sequence and also are not present in normal samples. VCMM [49] uses a simple multinomial model to compare the probability of the variant being a real variant with the probability of it being a sequencing error. Previous studies [93] have shown that MuTect is extremely conservative in its approach to identify somatic variants and has a high precision at the cost of identifying real somatic variants as germline. VCMM, on the other hand, tends to be liberal in its approach, since it does not take into consideration the corresponding normal sample and hence identifies many germline variants

as somatic. Overall VCMM has the highest sensitivity, while MuTect has the highest precision for detecting somatic variants [93].

The recent ICGC-TCGA Dream Mutation Calling challenge used crowd-sourcing to improve identification of somatic variants on one platform [94]. Different participating groups for the challenge have shown that ensemble approaches that integrate calling results from multiple somatic variant callers improve the identification of somatic variants over individual callers. For example, Kim et al. developed a statistical model that combines multiple callers [95]. SomaticSeq [96] uses Adaptive Boosting model, and CAKE [97] uses majority voting to classify a variant as somatic. It is worth noting that the aforementioned ensemble approaches have all been developed for identifying somatic variants from a single platform (i.e., a single data type, e.g., whole exome).

The low cost and high sequencing coverage associated with exome sequencing platform when compared to the whole genome sequencing platform, has led to more exome sequencing than genome sequencing. However, a recent study has shown that the whole genome sequencing accurately identifies more germline variants than whole exome sequencing in the exon regions [98]. Even though the study is for germline variant identification, it nevertheless suggests that relying only on exome data for variant calling may miss many real variants. In fact, our previous study [93] has shown that the concordance between somatic variants identified by whole exome data and by exonic regions of whole genome data is at most only 11%, and almost 90% of the somatic variants are called by only one of these two platforms, suggesting that there is much

room to explore with the two types of data for somatic variant calling. This phenomenon is also seen in germline variants, where concordance between whole exome and exonic regions of whole genome is low, approximately ~53% [89]. The important question in these cases is which platform analysis should be trusted or rather how can we make the best use of the two types of data for better somatic variant calling whenever both data types are available?

To address this question, we develop a framework that integrates the whole exome data and the whole genome data for somatic variant calling. Using two commonly used somatic variant callers, MuTect and VCMM, the former shown to have the highest precision and the latter the highest sensitivity [93] to call somatic variants on both exome and genome data, we then extracted 108 features from the calling outputs of the two programs and used them as input to the machine learning algorithm to identify somatic variants.

3.3 Results

Somatic mutations were generated on chromosome 1 of the individual "A0BW" from TCGA [99]. The whole genome had 30X coverage and the whole exome had 150X coverage. Somatic mutations were generated using BAMSurgeon [94] at different depths and different allele fractions on whole genome and whole exome platforms (see Methods 3.5.1 for details).

3.3.1 Number of somatic variants identified by callers individually

The somatic mutations that are generated by BAMSurgeon and that are also called by the somatic variant caller are considered as true positives. The false negatives (FN) are the simulated

variants that were not called by the variant caller. The false positives (FP) are the variants called by the variant caller but are not simulated variants. All the sites that are not simulated variants and are not called by the variant caller are true negatives (TN). Sensitivity was calculated using the formula TP/(TP+FN). Precision was calculated using the formula TP/(TP+FP). The F1 score is calculated using the formula (2*precision*recall)/(precision+recall). Table 3.1 shows the number of somatic variants identified by different somatic variant callers for the simulated whole genome and whole exome samples. Out of all the methods, VCMM has the highest sensitivity of 0.78 but very low precision, while MuTect has the highest precision value of 0.88 (similar to SomaticSniper). To improve the effectiveness of our framework, we decided to combine the results from MuTect and those from VCMM.

3.3.2 Results from different machine learning models

Figure 3.1 shows the results of a 10-fold cross-validation procedure using different machine learning classification models to identify true somatic variants. All classification algorithms used for this study were implemented in the Waikato Environment for Knowledge Analysis (WEKA). There were altogether 570,575 positions that were considered by MuTect and VCMM for somatic variant calling. The training set was built by randomly selecting 10,000 positions from the 570,575 positions. MuTect and VCMM were applied to the 10,000 sites. From the results of these two callers, 108 features were collected and used for machine learning. Comparison of different classifiers (only results of SMO, J48, MultiBoostAB, and DecisionTable with F1 scores higher than 0.90 are shown for brevity) shows that J48 has the highest F1 score of 0.968, and therefore was chosen as the classifier in the current study for further analysis.

3.3.3 Reason for integration of multiple tools and multiple data sets

Figure 3.2 shows the results of using J48 when compared to simple union and simple intersection of variants identified from only MuTect, SomaticSniper, VarScan2, and VCMM. Using J48 gives a sensitivity, precision, and F1 score of 0.94, 0.99 and 0.968, respectively. A union of somatic variants using MuTect, SomaticSniper, VarScan2, and VCMM gives an F1 score of 0.84, 0.84, 0.83, and 0.002 respectively; whereas a simple intersection of somatic variants produces F1-scores of 0.72, 0.72, 0.69 and 0.02 respectively using MuTect, SomaticSniper, VarScan2, and VCMM. This shows that our ensemble method which integrates multiple tools is better than individual callers in both sensitivity and precision. Figure 3.3 shows a distribution of the simulated variants that were detected by J48 (i.e., true positives) and the simulated variants that were missed by J48 (i.e., false negatives) across the coverage depth and allele fractions of the whole genome and the whole exome. Most somatic variants that J48 could not call as somatic had a low allele fraction in the genome and a low exome depth.

Figure 3.4 shows why combining whole genome and whole exome variants is better than calling variants from only one platform. Only somatic variants that were simulated on both whole genome and whole exome were considered for this part of the analysis. As seen in Figure 3.4, if only whole exome was considered the maximum sensitivity obtained was 0.59 (SomaticSniper) while if only whole genome was considered the maximum sensitivity obtained was 0.83 (VCMM). Using simple union of both whole genome and whole exome variant calling gives a highest sensitivity of 0.93 (VCMM). Using J48 to integrate somatic variants from both whole genome and whole exome platforms helps achieve a sensitivity of 0.95. This clearly

shows that integrating data sets from multiple platforms is better than just considering variants from one particular platform.

3.3.4 Results for cross-contamination of normal samples

It is known that Mutect imposes a heavy penalty on variants that are also present in the normal samples to prevent cross-contamination. However, this practice can also miss the true somatic variants and thus increase the number of false negatives. Therefore, it is necessary to study the effect of cross contamination on variant calling for the integrated caller. The normal samples were hence contaminated with allelic reads from the tumor samples at different percentages, i.e., 2.5%, 5%, 7.5%, and 10%. It is observed that as the degree of contamination increases, the F1 score decreases. Normal samples whose reads were replaced with allelic reads from the tumor samples at 2.5%, 5%, 7.5%, and 10% had 10-fold cross-validation scores of F1 scores of 0.96, 0.95, 0.93, 0.91 (Sensitivity of 0.93, 0.92, 0.89, and 0.86 and Precision of 0.99, 0.98, 0.97, and 0.97) respectively.

3.3.5 Comparison with similar tools

Since our framework is the first of its kind in integrating both multiple tools and multiple platforms, we could not compare our framework to another software in the same domain. We compared J48 to SomaticSeq [96], a tool that uses machine learning (Adaptive Boosting model implemented in R) to integrate somatic variant calling from multiple tools (MuTect, JointSNVMix2, SomaticSniper, VarDict, and VarScan2) from only whole genome or only whole exome platform to identify somatic variants. We applied SomaticSeq using the default trained

model built from high quality synthetic data. To make a fair comparison, we only compared somatic variants simulated in the non-exonic regions of the genome. Figure 3.5 shows that J48 performs better than SomaticSeq, achieving a sensitivity of 0.86 against SomaticSeq's 0.76.

3.3.6 Real data validation

For real data, we do not know the true somatic variants. Hence, variants that were identified by at least two somatic variant callers out of four somatic variant callers, i.e., MuTect, SomaticSniper, VarScan2, and VCMM, and from at least two platforms out of the three platforms (i.e., WGS, WXS and validation BAM files) available on TCGA were treated as true somatic variants (positives). Variants that were identified by only one platform or by only one somatic variant caller but were covered by another platform by a depth of at least 10X were considered to be true negatives. Our ensemble method based on a 10-fold cross validation on real data gives sensitivity, precision, and F1 score of 0.85, 0.80, and 0.83 respectively for A15K.

One of the ways to verify that the ensemble method works better than using only individual somatic variant callers is to show that the ensemble method can find somatic point mutations present in the whole genome but not in the whole exome despite these regions being covered by the whole exome sequencing. To make sure that the regions were true somatic point mutations, we searched for regions that were called by more than two somatic variant callers (MuTect, SomaticSniper, VarScan2, and VCMM) for the whole genome and also were present in the COSMIC database. We found 50 such positions that were identified by more than two somatic variant callers in the whole genome of individual "A15K" and were present in the

COSMIC database but were not called by any of the somatic variant callers from the whole exome BAM file even though they were covered by the whole exome sequencing. This shows our ensemble approach identifies variants that would have been missed if an intersection of the whole genome and whole exome somatic variants was considered as the method for the identification of somatic point mutations.

3.3.7 Robustness of the ensemble method

We also assessed the performance of our ensemble method using the same training set mentioned above but another test data set to examine the robustness of the trained model. To do this we produced a test data set from another individual (A15E) from TCGA. The test data set was produced using the MuTect, SomaticSniper, VarScan2, and VCMM results from individual A15E using the procedure mentioned above, i.e., variants from at least 2 callers on at least 2 platforms are positives while variants identified on only one platform by only one somatic variant caller and covered by a depth of at least 10X are negatives. We used the training set combined from A15K, and the data sets from A15E as test sets to check for the performance using the training set. This gave an F1 score of 0.567 (Figure 3.6). To increase the robustness of the training set, the AOBW data set was added to A15K. This was tested on A15E, which gives an F1 score of 0.617. We then added A152 data set to the combination of A15K and A0BW. The combination of A15K-A0BW-A152 was tested on A15E, which gives an F1 score of 0.681.

3.4 Discussion

In somatic variant calling, an ideal case scenario would be if a variant is identified on both whole genome and whole exome platforms as this would also be a strong corroboration of a somatic variant at that particular position. However, in reality, there could be variants that are identified in the whole genome data but not identified in the whole exome and vice-versa. In fact, it has been shown that germline variants called on both platforms account for only 53% of the total variants called on the two platforms [89], and for somatic variants, the proportion of variants that are called on both platforms is even smaller, only about 11% [93]. One of the reasons for the low consensus could be the region not being sequenced at a high enough coverage in either of the platforms. If the allele fraction is too low, the allele may not have enough coverage in one of the platforms and thus may have been considered as a sequencing error. The region can be a low complexity region such as a repeat region and, hence, is difficult to identify. To address this issue and overcome the disagreement between different platforms, we present a framework that can be used to facilitate our aim to improve the identification of somatic variants.

Generally, either an intersection or union of called variants from whole genome and whole exome platforms is taken into account to identify variants for an individual. However, using a simple union of variants could lead to the incorporation of many false positives; using simple intersection could lead to the exclusion of many true positives, especially considering the observation that the concordance between exome and genome within the exonic regions did not exceed 11% for any of the commonly used somatic variant callers [93]. It has been shown that

more variants are identified from the exonic regions of the whole genome than the whole exome in the case of germline variants [100]. Our study corroborates this in the case of somatic variants (Figure 3.4). One of the approaches that could be considered is combining the whole genome and whole exome BAM files. The advantage of this approach is that the depth would increase at regions where the exome or genome had lower depth. In the regions where whole genome and whole exome have high depth, somatic variant callers would not identify variants correctly as regions with extremely high depth are considered as low complexity regions. For example, VCMM does not analyze regions with depth higher than 700 to identify somatic variants. The observation that integration of multiple variant calling tools is better than using a tool individually has also been shown before [101]. Hence, we provide a framework that uses an ensemble approach to incorporate variants from both whole exome and whole genome platforms using multiple somatic variant callers without adding too many false positives and missing too many true positives.

To resolve the disagreement between the variants detected by the two platforms, we developed an ensemble method that combines the outputs from MuTect and VCMM. The output file of MuTect using the "call_stats" options gives details of the variants in normal and tumor samples and gives reasons for why a variant was accepted as a somatic variant or why it was rejected. We focus on integrating MuTect and VCMM, since it has been shown that MuTect has a high precision while VCMM has a high sensitivity [93]. VCMM predicts 100 times the number of somatic variants that MuTect, VarScan2 or SomaticSniper predict and, thus, predicts many false positives. VCMM does not take the normal sample into consideration and hence predicts a

lot of somatic variants. According to our previous study, approximately 80% of the variants that VCMM predicts are germline variants. So it is important to restrict the number of somatic variants that VCMM predicts with the help of MuTect, which heavily penalizes the presence of variants in normal samples. On the other hand, since MuTect is extremely conservative in its approach, it is necessary to use VCMM to increase the identification of true somatic variants.

We used J48 in our ensemble approach to classify the variants identified by MuTect and VCMM as somatic or not. Decision tree J48 in WEKA is an iterative Dichotomiser 3 (ID3) implementation. Decision trees are very advantageous since they can handle missing values and many types of data including nominal, numeric, and textual data. Our input data to J48 was obtained from the output of MuTect and VCMM on the whole genome and whole exome data. The input data to J48 includes much numeric and textual data. Attributes such as base quality and mapping quality are numerical, while attributes such as dbSNP and COSMIC are textual. If a variant was identified by MuTect but not identified by VCMM or vice-versa or if it is identified on only one of the platforms out of whole genome and whole exome, the attributes could have much missing data. A decision tree uses information gain for attribute selection. Information gain assigns an importance to each attribute by giving a cutoff value to each attribute to split the node into two leaves. We used all the 108 features, i.e., all the information collected from MuTect and VCMM output files. The 108 features include base quality, mapping quality, indel score, SNP quality, allele fraction, coverage of normal and tumor samples, and presence of the position in dbSNP or COSMIC database. These features were selected, because most somatic variant callers

use different and arbitrary cutoffs for the features. We let the machine learning algorithm decide a cutoff and determine whether the variants are truly somatic.

We show that the ensemble approach gives a high F1 score on both simulated and real data (Figure 3.2 and Figure 3.6). Using an ensemble approach reduces the number of false positives and hence the precision values of the ensemble method increases compared to precision values of individual callers (Figure 3.2). The sensitivity values are limited by the callers that are used in the ensemble method because the variants that are not identified by any method individually cannot be identified by the ensemble approach either. Intuitively, it makes sense to predict that variants that would not be detected either had a low allelic fraction or a low tumor sequencing depth in the whole genome or whole exome platforms, which is seen in our study (Figure 3.3). We also compared our framework to SomaticSeq, also a machine learning framework that combines results from multiple variant calling tools to identify variants from whole genome or whole exome. Our framework performs better than SomaticSeq to identify somatic variants on one platform (Figure 3.5). This shows that our framework of combining Mutect and VCMM can also be used effectively to identify somatic variants accurately from only one platform. Although we demonstrated our framework using VCMM and MuTect, we believe that our framework can also be applied to other tools.

In a clinical set up, it is still uncommon to obtain both normal and tumor samples, because of the costs associated with sequencing normal and tumor tissue samples. Identifying somatic variants from only tumor samples is difficult, because it would be challenging to

differentiate between germline variants, somatic variants, and sequencing errors. VCMM does not take into account the normal samples while identifying somatic variants. MuTect can also be used to identify somatic variants without the normal sample but this results in lower accuracy. A future direction for this work would be to improve identification of somatic variants using variants identified from only tumor samples.

3.5 Methods

3.5.1 Generating simulated data set

Simulated data sets for the whole exome and whole genome platforms were generated using BAMSurgeon. BAMSurgeon can add somatic variants to genome and exome platforms by adding mutations to particular sites at specified allele fraction. Somatic variants were simulated on chromosome 1 of the individual "A0BW" from TCGA. Real exome data usually have different coverage depths depending on the particular experiments. Therefore, 100 somatic variants were generated with coverage depths <=8x, <=14x, <=200x, <=500x, <=800x, and >800x for the whole exome data. Therefore, to closely reflect the observation in real data, 100 variants were also simulated in regions that were covered by the whole genome but not by the whole exome. Since difference in coverage depth between normal and tumor samples is also common in real data, to closely mimic the real data, a coverage difference of a maximum of 50% was set between the normal and tumor samples. The parameter *coverdiff* in BAMSurgeon is utilized to simulate the coverage depth difference between tumor and normal samples. Note that the difference in coverage between normal and tumor, the difference in depth and allele fraction in whole exome and genome is a representation of real data. Figure 3.7 shows the distribution of

simulated somatic point mutations across different allele fractions and coverage depths on whole genome and whole exome platforms.

3.5.2 Building training and test sets for simulated data

The simulated variants are considered as true somatic variants (positives). Variants that were heterozygous in the normal sample or germline variants and other sites detected as variants by Mutect and VCMM were identified as not somatic (negatives). The germline variants in the sample were identified using GATK HaplotypeCaller, since it has been shown to identify germline variants with high precision [87]. Output results from the "call_stats" option of Mutect and VCMM were then combined and the resulting file contains variants accepted as somatic variants by Mutect, variants rejected by Mutect due to various reasons, and variants called by VCMM. Altogether 108 features such as coverage depth, mapping quality, base quality, indel score, presence in dbSNP or COSMIC database, and number of bases in positive or negative strand were used to build the training set. The list of the 108 features as ranked by the InfoGain algorithm is shown in supplementary Table S3.1. A custom code used to build this features list is available upon request.

3.5.3 Models used to identify somatic variants

A number of classification tools were tested to identify the most suitable model that can be used to identify somatic variants. The classification algorithms that were tested and had an F1 score more than 0.90 were "J48", "SM0", "DecisionTable" and "MultiBoostAB". These algorithms were implemented as a part of the WEKA suite [102]. We compared 45 classifiers

with their default parameters and the aforementioned four classifiers had an F1-score higher than 0.90 (as high as 0.98 on simulated data).

J48 was developed by Ross Quinlan to generate an iterative decision tree. J48 builds decision trees using the concept of information gain. J48 identifies the feature that helps classify the training set the best. This feature with the highest normalized information gain is chosen to classify the training set. This process is continued using all the attributes until all the records in the training set can be classified accurately. J48 can also classify features that have continuous value. The training set with continuous values is classified such that the gain is maximized.

SMO (Sequential Minimal Optimizer) is an implementation of a support vector machine. SMO partitions the training data into smaller optimization sub-problems that can be solved analytically using two Lagrange multipliers. Splitting the data into smaller problems helps save space and memory. All missing values in the data are replaced with normalized values and nominal attributes are transformed into binary ones.

A DecisionTable is an elegant if-else condition table. Given a set of features, the best feature based on which the data can be classified is chosen. Consecutively features are chosen one by one based on how well they can classify the data. Each record in the training data consists of a value for each feature. The records with values within a particular range in the different features, are classified in one class.

MultiBoostAB uses adaptive boosting to classify the data. Adaptive boosting builds upon a weak classifier. In the case of adaptive boosting, data is reclassified after learning from a weak classifier, which is J48 in WEKA. Data is reclassified by reweighing the data. Data that is misclassified gains weight while data that is classified correctly loses weight. The reweighed misclassified data is then reclassified until the data is classified correctly.

3.5.4 Building training and test sets for real data

To examine the performance of the machine learning model, real data was also used. The DNA of the individual "A15K" has been sequenced three different times on different platforms, i.e., whole genome, whole exome, and validation BAM files (available on TCGA). Since we do not know the actual somatic variants in an individual with tumor, variants were considered as positives (i.e., truly somatic) if they were called as somatic by at least two methods (out of Mutect, SomaticSniper, Varscan2, and VCMM) on at least two platforms (whole genome, whole exome, and validation BAM files). Using variants from multiple platforms by multiple callers as true variants in real data has been used before by different studies [103]. Variants were considered as negatives (i.e., not somatic) if they were identified as somatic by only one method on one platform and had a coverage depth of at least 10X in the normal and tumor samples on the other two platforms. For example, if a variant was identified by only one caller on whole genome but was not identified by any other caller on whole exome or validation BAM files, the variant was considered as negative. This 10X coverage was selected because most methods require a minimum depth of 8-12X [90] to identify somatic variants.

Another test data was generated to validate the training set. The individual "A15E" was used for this purpose. The methodology described above to identify true and false somatic variants was used to build the test data, i.e., variants identified as somatic by multiple callers from multiple platforms are considered as positives while variants identified by only one somatic variant caller on one platform but not identified by any other variant caller on the other two platforms are considered as negatives.

To build a robust training set for real data, we combined variants from three real data sets, i.e., A15K, A0BW and A152. There were 2419 true positive examples and 26637 negative examples in the training set. We built another robust training set for real data, because it would be difficult to encompass the distribution in base quality, mapping quality, allele fraction, low coverage of alleles, nearby indels, and nearby repeat regions in simulated data.

3.6 Conclusions

We developed a framework that integrates somatic point mutations called by two somatic variant callers MuTect and VCMM from two platforms i.e., whole genome and whole exome. We used 108 attributes from the MuTect and VCMM outputs as input to decision tree classifier J48 to classify the variants from MuTect and VCMM as truly somatic or not. This ensemble method works better than using individual calling methods, or using the simple union or intersection of variants called by the methods. Using this ensemble approach only on whole genome or only whole exome platforms also works better than using only one method individually, showing that the approach is promising.

3.7 Abbreviations

COSMIC: Catalogue Of Somatic Mutations In Cancer

FNs: False Negatives

FPs: False Positives

ID3: Iterative Dichotomiser 3

TCGA: The Cancer Genome Atlas

TNs: True Negatives

TPs: True Positives

VCMM: Variant Caller with Multinomial probabilistic Model

WEKA: Waikato Environment for Knowledge Analysis

WGS: Whole Genome Sequence

WXS: Whole Exome Sequence

Figures

Figure 3.1: F1 score for different machine learning models.

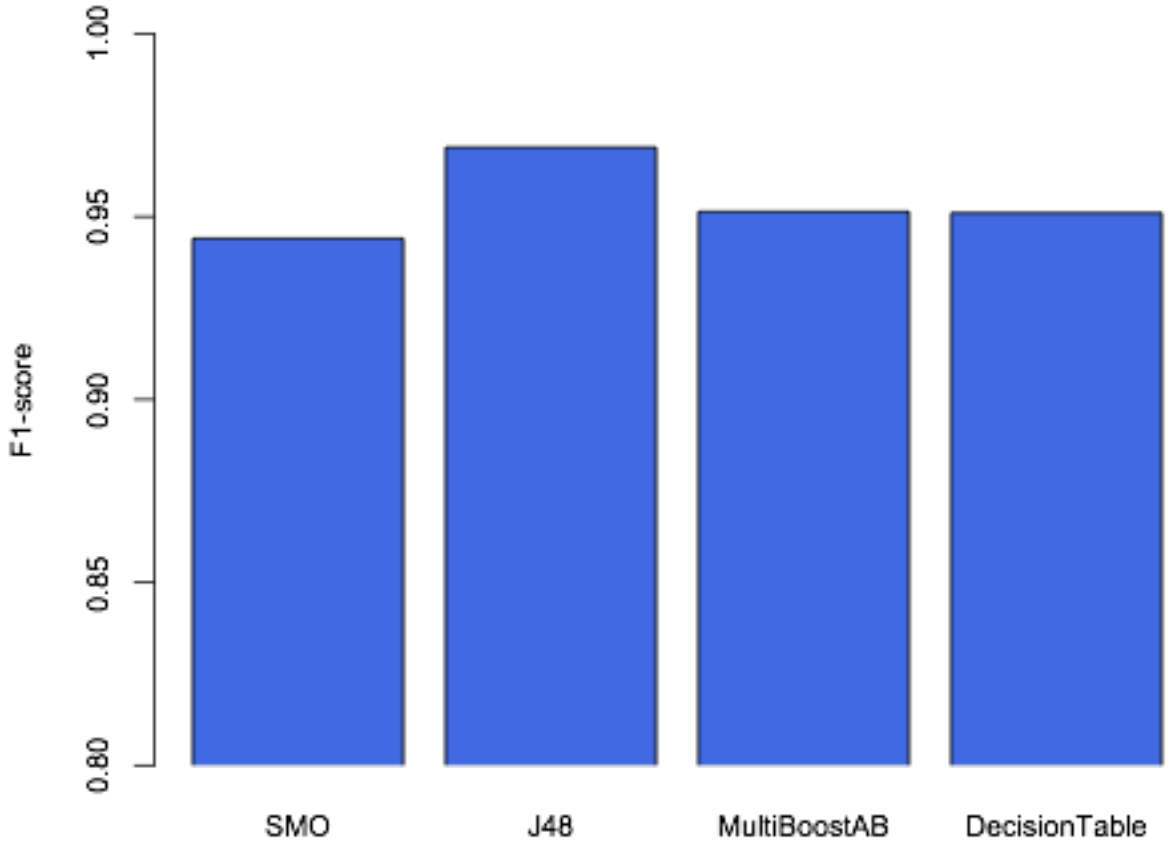


Figure 3.2: Sensitivity, precision and F1 score with MuTect, SomaticSniper, VarScan2 VCMM and J48.

'u' indicates union and 'i' indicates intersection results.

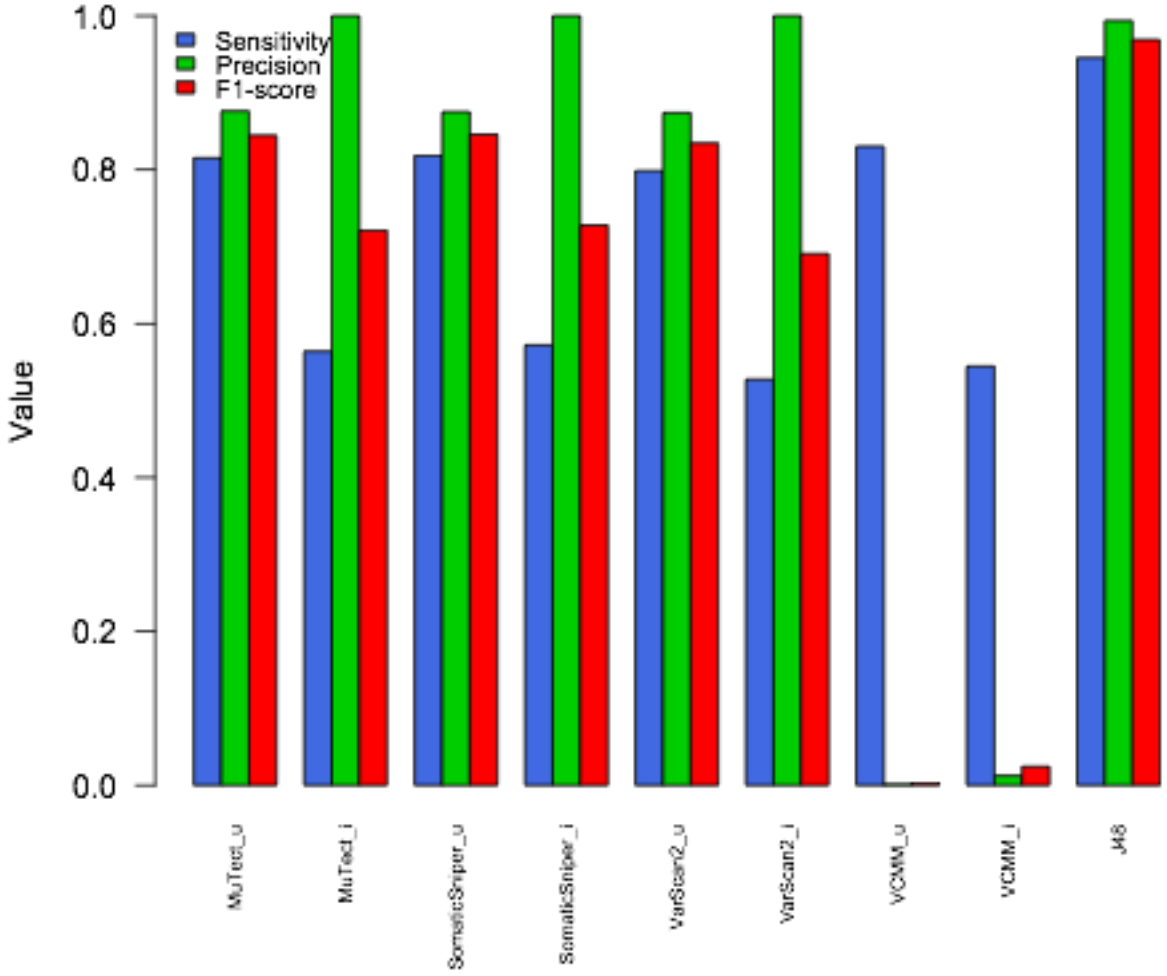


Figure 3.3: Distribution of the true positives and false negatives across the depth and allele fractions of whole genome and whole exome.

Red rhombus depict true positives and blue dots depict false negatives.

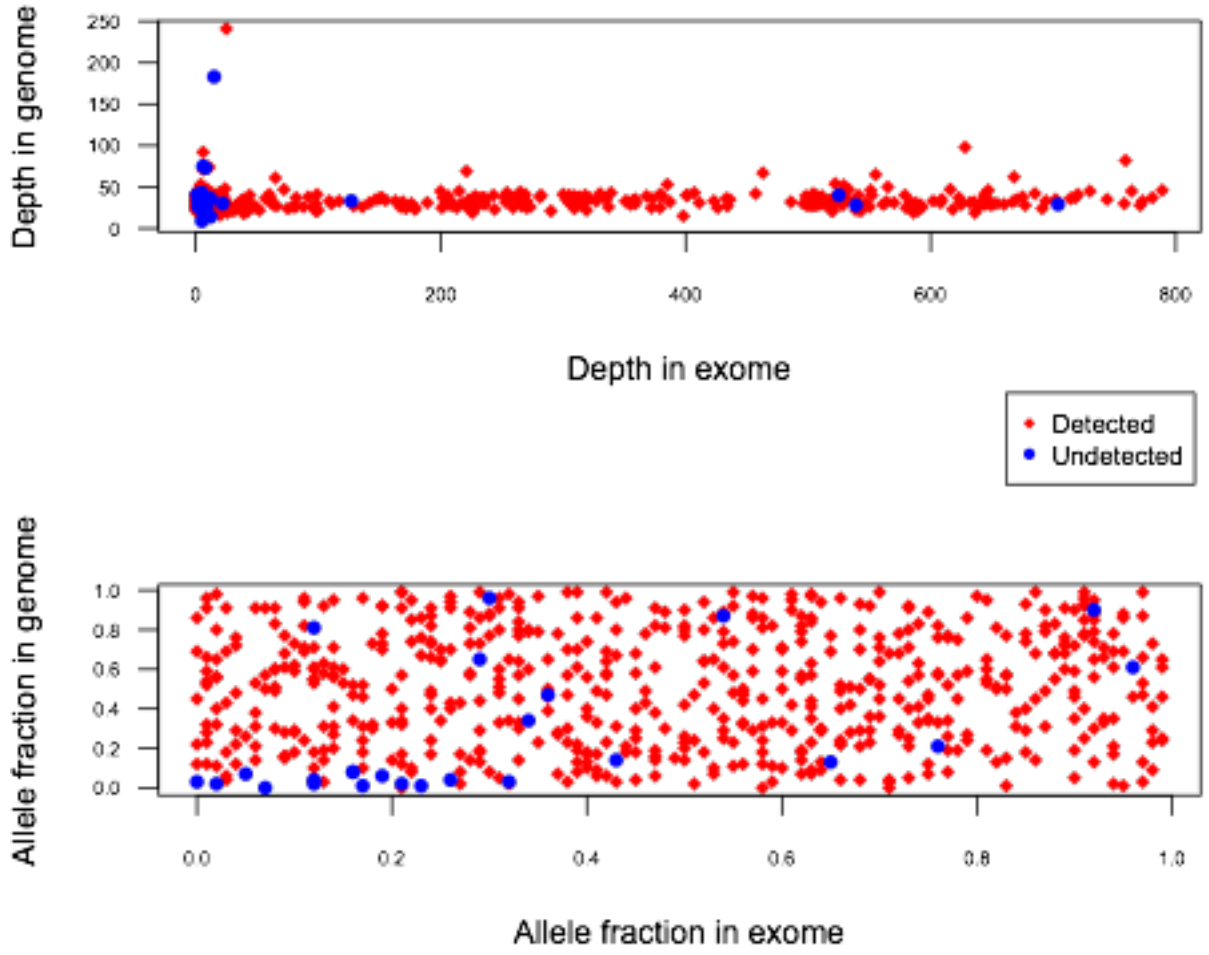


Figure 3.4: Performance comparison of somatic variant identification for single platform, i.e., whole genome (WGS) or whole exome (WXS) versus ensemble method.

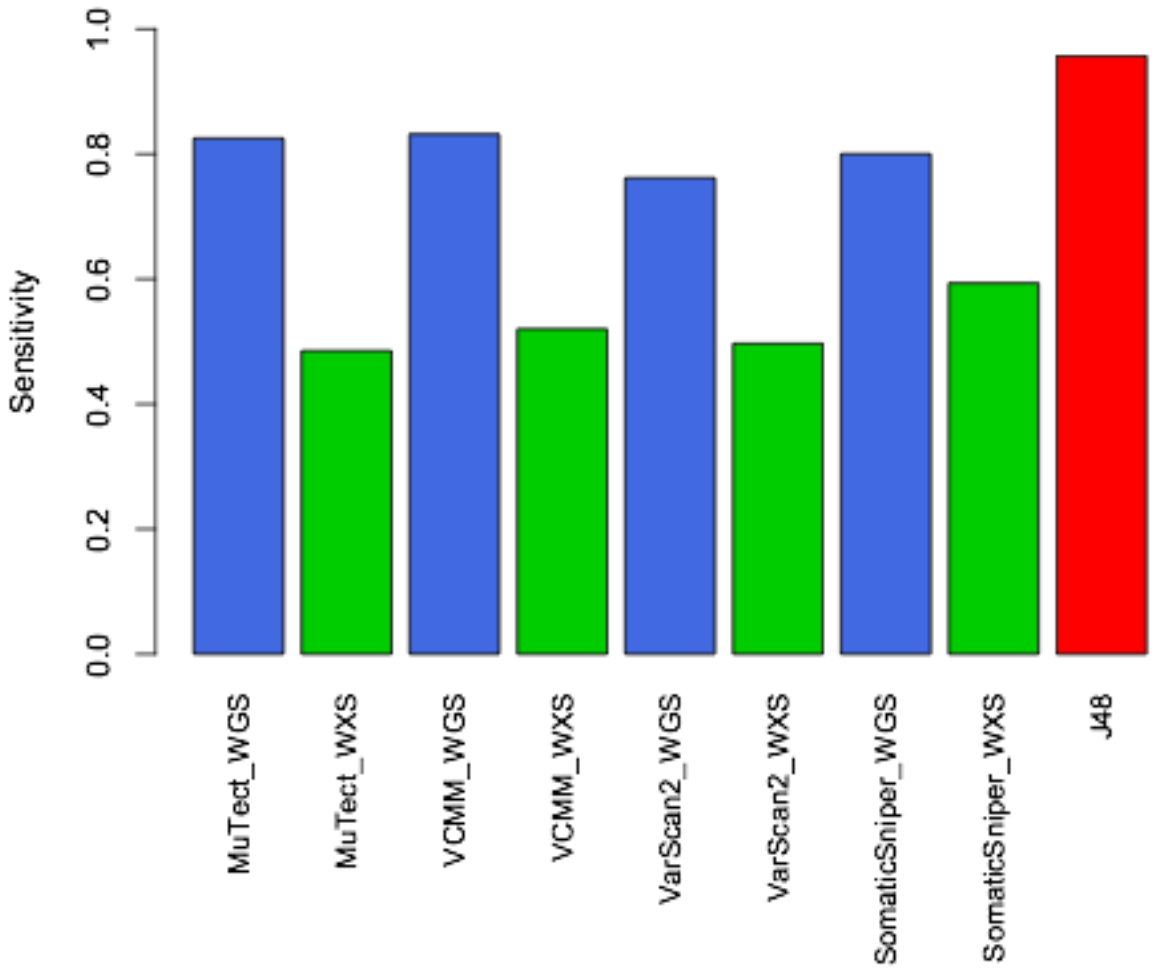


Figure 3.5: Performance comparison of SomaticSeq versus our ensemble method on only whole genome platform.

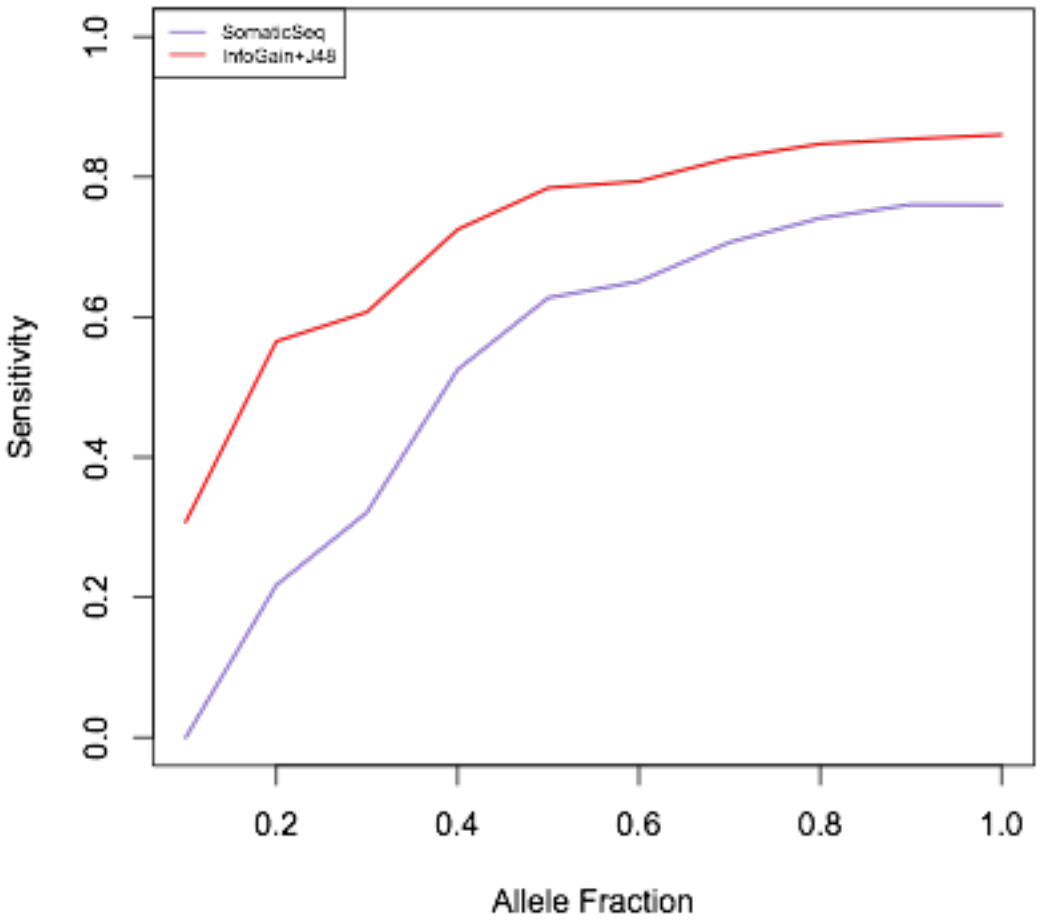


Figure 3.6: Sensitivity, precision and F1 scores based on different training sets

(i) A15K, (ii) combination of A15K-A0BW, and (iii) combination of A15K-A0BW-A152, using A15E as the test set.

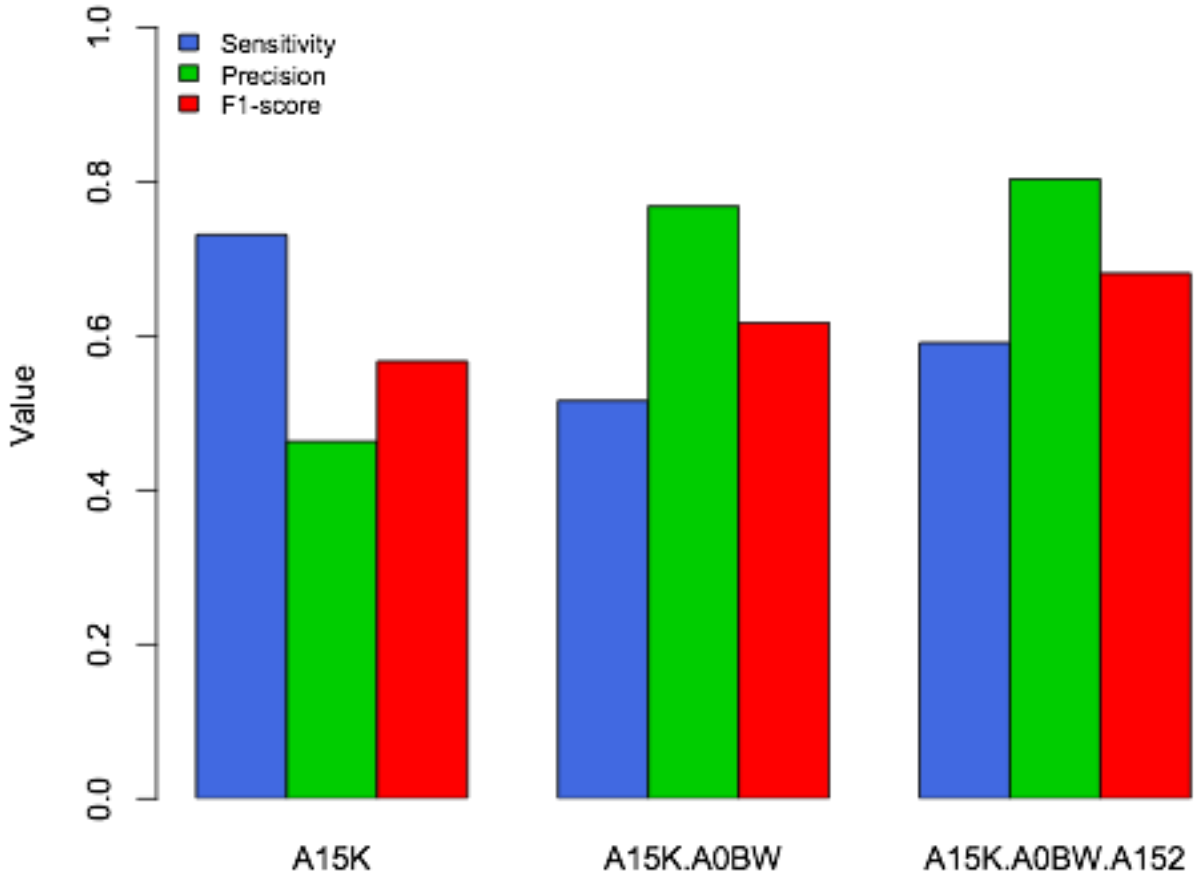
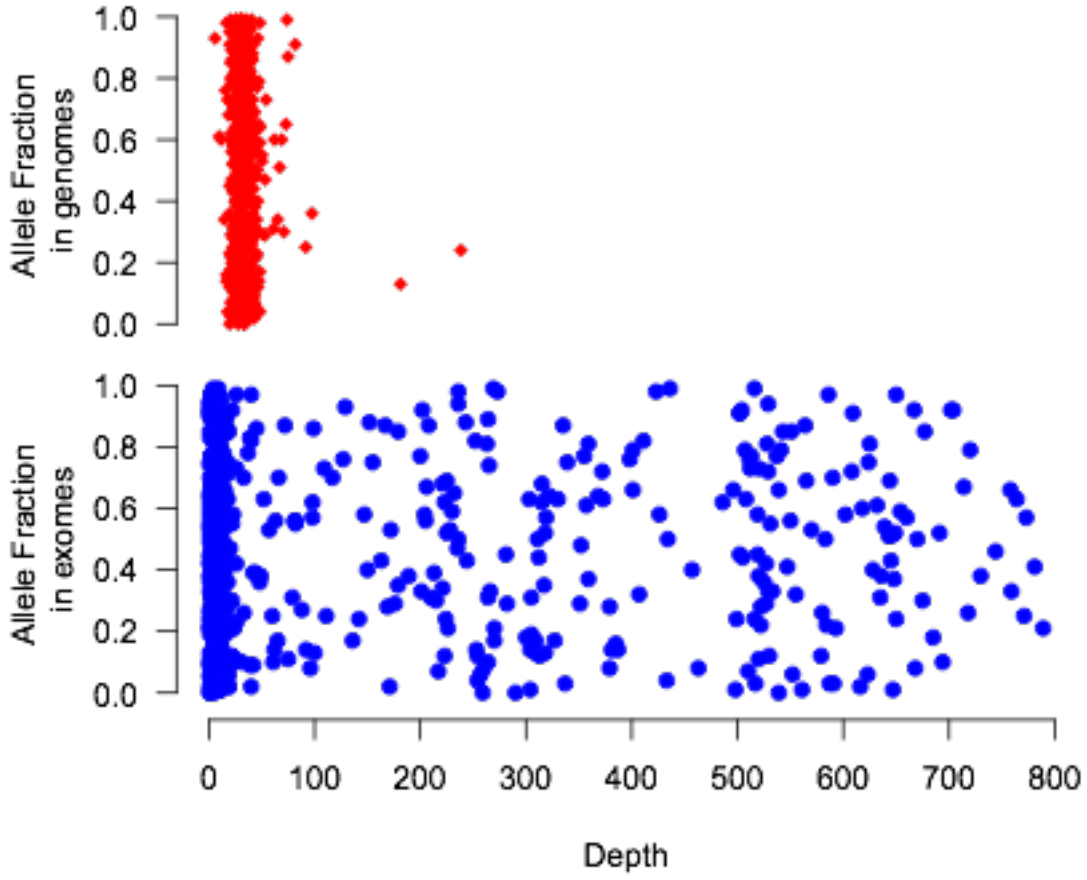


Figure 3.7: Distribution of simulated somatic point mutations across different allele fractions and depths on whole genome and whole exome platforms.



Tables

Table 3.1: The number of somatic variants called by four methods.

[1]Number of true positives, computed as the simple union of the number of variants called by each method out of the 700 simulated somatic variants for the two platform data. [2]Total number of somatic variants called by each method on the whole exome and whole genome data sets.

Method	# of True positives[1]	# of Somatic Variants[2]	Sensitivity	Precision
MuTect	538	614	0.77	0.88
SomaticSniper	540	617	0.77	0.88
VarScan2	527	603	0.75	0.87
VCMM	548	399,491	0.78	0.0014

Supplementary Tables

Table S3.1: The number of features as ranked by InfoGain algorithm.

*G indicates genome, X indicates exome, M indicates MuTect, V indicates VCMM. See [84] for further details on the parameters.

Rank	Parameter	Description*
1	tot_depthG	depth (G,V)
2	refG	Reference allele (G,V)
3	altG	Alternate allele(G,V)
4	p_allG	log10(p-allele)(G,V)
5	p_errG	log10(p-error) (G,V)
6	snp_qualG	SNP-quality (G,V)
7	ind_filterG	neighbor-indel-filter (G,V)
8	snp_filterG	neighbor-snp-filter (G,V)
9	tot_depthX	depth (X,V)
10	refX	Reference allele (X,V)
11	altX	Alternate allele(X,V)
12	p_allX	log10(p-allele)(X,V)
13	p_errX	log10(p-error) (X,V)
14	snp_qualX	SNP-quality (X,V)
15	ind_filterX	neighbor-indel-filter (X,V)
16	snp_filterX	neighbor-snp-filter (X,V)
17	scoreX	score (X,M)
18	dbsnp_siteX	dbsnp_site (X,M)
19	coveredX	covered (X,M)
20	powerX	power (X,M)

Rank	Parameter	Description*
21	tumor_powerX	tumor_power (X,M)
22	normal_powerX	normal_power (X,M)
23	normal_power_nspX	normal_power_nsp (X,M)
24	normal_power_wspX	normal_power_wsp (X,M)
25	total_readsX	total_reads (X,M)
26	map_Q0_readsX	map_Q0_reads (X,M)
27	init_t_lodX	init_t_lod (X,M)
28	t_lod_fstarX	t_lod_fstar (X,M)
29	t_lod_fstar_forX	t_lod_fstar_for (X,M)
30	t_lod_fstar_revX	t_lod_fstar_rev (X,M)
31	tumor_fX	tumor_f (X,M)
32	contaminant_fracX	contaminant_frac (X,M)
33	contaminant_lodX	contaminant_lod (X,M)
34	t_q20_countX	t_q20_count (X,M)
35	t_ref_countX	t_ref_count (X,M)
36	t_alt_countX	t_alt_count (X,M)
37	t_ref_sumX	t_ref_sum (X,M)
38	t_alt_sumX	t_alt_sum (X,M)
39	t_ref_max_mapqX	t_ref_max_mapq (X,M)
40	t_alt_max_mapqX	t_alt_max_mapq (X,M)
41	t_ins_countX	t_ins_count (X,M)
42	t_del_countX	t_del_count (X,M)
43	normal_best_gtX	normal_best_gt (X,M)
44	init_n_lodX	init_n_lod (X,M)
45	normal_fX	normal_f (X,M)
46	n_q20_countX	n_q20_count (X,M)

Rank	Parameter	Description*
47	n_ref_countX	n_ref_count (X,M)
48	n_alt_countX	n_alt_count (X,M)
49	n_ref_sumX	n_ref_sum (X,M)
50	n_alt_sumX	n_alt_sum (X,M)
51	Power_Pos_SBX	power_to_detect_positive_strand_artifact (X,M)
52	Power_Neg_SBX	power_to_detect_negative_strand_artifact (X,M)
53	SB1X	strand_bias_counts1 (X,M)
54	SB2X	strand_bias_counts2 (X,M)
55	SB3X	strand_bias_counts3 (X,M)
56	SB4X	strand_bias_counts4 (X,M)
57	AT_FmedX	tumor_alt_fpir_median (X,M)
58	AT_FmadX	tumor_alt_fpir_mad (X,M)
59	AT_RmedX	tumor_alt_rpir_median (X,M)
60	AT_RX	tumor_alt_rpir_mad (X,M)
61	NormX	observed_in_normals_count (G,M)
62	judgementX	judgement (G,M)
63	scoreG	score (G,M)
64	dbsnp_siteG	dbsnp_site (G,M)
65	coveredG	covered (G,M)
66	powerG	power (G,M)
67	tumor_powerG	tumor_power (G,M)
68	normal_powerG	normal_power (G,M)
69	normal_power_nspG	normal_power_nsp (G,M)
70	normal_power_wspG	normal_power_wsp (G,M)

Rank	Parameter	Description*
71	total_readsG	total_reads (G,M)
72	map_Q0_readsG	map_Q0_reads (G,M)
73	init_t_lodG	init_t_lod (G,M)
74	t_lod_fstarG	t_lod_fstar (G,M)
75	t_lod_fstar_forG	t_lod_fstar_forward (G,M)
76	t_lod_fstar_revG	t_lod_fstar_reverse (G,M)
77	tumor_fG	tumor_f (G,M)
78	contaminant_fracG	contaminant_fraction (G,M)
79	contaminant_lodG	contaminant_lod (G,M)
80	t_q20_countG	t_q20_count (G,M)
81	t_ref_countG	t_ref_count (G,M)
82	t_alt_countG	t_alt_count (G,M)
83	t_ref_sumG	t_ref_sum (G,M)
84	t_alt_sumG	t_alt_sum (G,M)
85	t_ref_max_mapqG	t_ref_max_mapq (G,M)
86	t_alt_max_mapqG	t_alt_max_mapq (G,M)
87	t_ins_countG	t_ins_count (G,M)
88	t_del_countG	t_del_count (G,M)
89	normal_best_gtG	normal_best_gt (G,M)
90	init_n_lodG	init_n_lod (G,M)
91	normal_fG	normal_f (G,M)
92	n_q20_countG	n_q20_count (G,M)
93	n_ref_countG	n_ref_count (G,M)
94	n_alt_countG	n_alt_count (G,M)
95	n_ref_sumG	n_ref_sum (G,M)
96	n_alt_sumG	n_alt_sum (G,M)

Rank	Parameter	Description*
97	Power_Pos_SBG	power_to_detect_positive_strand_artifact (G,M)
98	Power_Neg_SBG	power_to_detect_negative_strand_artifact (G,M)
99	SB1G	strand_bias_counts1 (G,M)
100	SB2G	strand_bias_counts2 (G,M)
101	SB3G	strand_bias_counts3 (G,M)
102	SB4G	strand_bias_counts4 (G,M)
103	AT_FmedG	tumor_alt_fpir_median (G,M)
104	AT_FmadG	tumor_alt_fpir_mad (G,M)
105	AT_RmedG	tumor_alt_rpir_median (G,M)
106	AT_RG	tumor_alt_rpir_mad (G,M)
107	NormG	observed_in_normals_count (G,M)
108	judgementG	judgement (G)

Figure S3.1: The tree built by J48 using the model training set i.e. A0BW-A15K-A152

```
judgementG = KEEP
|   init_t_lodG <= 5.089038
|   |   Power_Pos_SBG <= 0.801398: DISCARD (188.02/6.12)
|   |   Power_Pos_SBG > 0.801398
|   |   |   AT_FmadG <= 15.5
|   |   |   |   coveredX = COVERED: DISCARD (16.72/3.01)
|   |   |   |   coveredX = UNCOVERED
|   |   |   |   |   t_alt_sumG <= 195: DISCARD (2.25/0.0)
|   |   |   |   |   t_alt_sumG > 195: ACCEPT (3.27/0.27)
|   |   |   AT_FmadG > 15.5: ACCEPT (4.04/0.04)
|   init_t_lodG > 5.089038
|   |   n_alt_sumX <= 26: ACCEPT (806.46/7.65)
|   |   n_alt_sumX > 26
|   |   |   t_ref_max_mapqX <= 52: DISCARD (2.85/0.11)
|   |   |   t_ref_max_mapqX > 52: ACCEPT (23.63/4.01)
judgementG = REJECT
|   n_alt_sumG <= 202
|   |   total_readsG <= 1434
|   |   |   normal_fG <= 0.218014
|   |   |   |   snp_qualG <= 11.373
|   |   |   |   |   t_del_countG <= 5
|   |   |   |   |   |   tumor_fG <= 0.16129
|   |   |   |   |   |   |   snp_qualG <= 7.732: DISCARD (619.02/47.59)
|   |   |   |   |   |   |   snp_qualG > 7.732
|   |   |   |   |   |   |   |   tumor_fX <= 0.122222
|   |   |   |   |   |   |   |   |   map_Q0_readsG <= 9
|   |   |   |   |   |   |   |   |   |   normal_power_nspG <= 0.99649
|   |   |   |   |   |   |   |   |   |   |   t_ref_max_mapqG <= 48: DISCARD (2.48/0.36)
|   |   |   |   |   |   |   |   |   |   |   t_ref_max_mapqG > 48: ACCEPT (24.67/9.58)
|   |   |   |   |   |   |   |   |   |   normal_power_nspG > 0.99649
|   |   |   |   |   |   |   |   |   |   |   coveredX = COVERED: DISCARD (179.59/33.6)
|   |   |   |   |   |   |   |   |   |   |   coveredX = UNCOVERED
|   |   |   |   |   |   |   |   |   |   |   |   normal_power_nspX <= 0: ACCEPT
(7.32/1.15)
|   |   |   |   |   |   |   |   |   |   |   |   normal_power_nspX > 0: DISCARD
(8.67/1.76)
|   |   |   |   |   |   |   |   |   map_Q0_readsG > 9: DISCARD (35.59/0.18)
|   |   |   |   |   |   |   |   tumor_fX > 0.122222
|   |   |   |   |   |   |   |   |   SB4G <= 0: DISCARD (15.87/7.23)
|   |   |   |   |   |   |   |   |   SB4G > 0: ACCEPT (20.25/2.7)
|   |   |   |   |   |   tumor_fG > 0.16129
|   |   |   |   |   |   |   total_readsX <= 35
|   |   |   |   |   |   |   |   powerX <= 0.099721
|   |   |   |   |   |   |   |   |   dbsnp_siteG = NOVEL
|   |   |   |   |   |   |   |   |   |   snp_qualG <= 6.963: DISCARD (2.53/0.18)
|   |   |   |   |   |   |   |   |   |   snp_qualG > 6.963: ACCEPT (8.89/0.71)
|   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP: ACCEPT (21.3/2.66)
|   |   |   |   |   |   |   |   |   dbsnp_siteG = COSMIC: ACCEPT (0.0)
|   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP+COSMIC: DISCARD (0.63/0.0)
|   |   |   |   |   |   |   |   powerX > 0.099721: DISCARD (3.87/1.13)
|   |   |   |   |   |   |   total_readsX > 35
|   |   |   |   |   |   |   |   judgementX = KEEP: ACCEPT (12.43/3.9)
|   |   |   |   |   |   |   |   judgementX = REJECT
|   |   |   |   |   |   |   |   |   dbsnp_siteG = NOVEL
|   |   |   |   |   |   |   |   |   |   coveredG = COVERED: DISCARD (24.16/3.75)
|   |   |   |   |   |   |   |   |   |   coveredG = UNCOVERED
|   |   |   |   |   |   |   |   |   |   |   t_del_countG <= 0: DISCARD (4.93/1.68)
|   |   |   |   |   |   |   |   |   |   |   t_del_countG > 0: ACCEPT (2.33/0.03)
|   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP
|   |   |   |   |   |   |   |   |   |   snp_qualG <= 4.786: DISCARD (18.64/0.83)
```

```
|   |   |   |   |   |   |   |   |   |   snp_qualG > 4.786
|   |   |   |   |   |   |   |   |   |   |   tumor_powerG <= 0.851984
|   |   |   |   |   |   |   |   |   |   |   |   n_ref_countG <= 4: DISCARD (7.01/0.0)
|   |   |   |   |   |   |   |   |   |   |   |   n_ref_countG > 4
|   |   |   |   |   |   |   |   |   |   |   |   |   SB3G <= 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   tumor_powerG <= 0.260965:
ACCEPT (5.43/1.99)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   tumor_powerG > 0.260965:
DISCARD (7.51/0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   SB3G > 0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   coveredX = COVERED: ACCEPT
(32.27/9.63)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   coveredX = UNCOVERED: DISCARD
(3.19/0.67)
|   |   |   |   |   |   |   |   |   |   |   tumor_powerG > 0.851984: DISCARD
(10.9/0.44)
|   |   |   |   |   |   |   |   |   |   dbsnp_siteG = COSMIC: DISCARD (7.77/3.0)
|   |   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP+COSMIC: DISCARD (16.06/0.57)
|   |   |   |   |   t_del_countG > 5
|   |   |   |   |   |   total_readsG <= 47
|   |   |   |   |   |   |   normal_fX <= 0.140845
|   |   |   |   |   |   |   |   t_alt_max_mapqG <= 14
|   |   |   |   |   |   |   |   |   n_alt_sumG <= 33: DISCARD (4.42/1.0)
|   |   |   |   |   |   |   |   |   n_alt_sumG > 33: ACCEPT (3.72/0.03)
|   |   |   |   |   |   |   |   t_alt_max_mapqG > 14
|   |   |   |   |   |   |   |   |   snp_qualG <= 4.281: DISCARD (5.08/1.36)
|   |   |   |   |   |   |   |   |   snp_qualG > 4.281: ACCEPT (74.57/5.75)
|   |   |   |   |   |   |   normal_fX > 0.140845
|   |   |   |   |   |   |   |   SB3X <= 1: ACCEPT (6.27/1.96)
|   |   |   |   |   |   |   |   SB3X > 1: DISCARD (10.01/2.16)
|   |   |   |   |   |   total_readsG > 47
|   |   |   |   |   |   |   t_alt_countG <= 6
|   |   |   |   |   |   |   |   refG <= 32
|   |   |   |   |   |   |   |   |   altG <= 5
|   |   |   |   |   |   |   |   |   |   n_q20_countG <= 30
|   |   |   |   |   |   |   |   |   |   |   SB1G <= 17: DISCARD (19.16/1.02)
|   |   |   |   |   |   |   |   |   |   |   SB1G > 17: ACCEPT (2.27/0.26)
|   |   |   |   |   |   |   |   |   |   n_q20_countG > 30: ACCEPT (2.27/0.02)
|   |   |   |   |   |   |   |   |   altG > 5: ACCEPT (8.34/1.25)
|   |   |   |   |   |   |   |   refG > 32: DISCARD (21.64/0.24)
|   |   |   |   |   |   |   t_alt_countG > 6: ACCEPT (3.03/0.03)
|   |   |   |   snp_qualG > 11.373
|   |   |   |   |   init_n_lodG <= -5.406265
|   |   |   |   |   |   map_Q0_readsG <= 3
|   |   |   |   |   |   |   t_ref_countX <= 23
|   |   |   |   |   |   |   |   normal_fX <= 0.135593: ACCEPT (30.14/4.11)
|   |   |   |   |   |   |   |   normal_fX > 0.135593
|   |   |   |   |   |   |   |   |   dbsnp_siteG = NOVEL
|   |   |   |   |   |   |   |   |   |   t_ins_countG <= 1: ACCEPT (8.28/1.59)
|   |   |   |   |   |   |   |   |   |   t_ins_countG > 1: DISCARD (3.03/0.0)
|   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP
|   |   |   |   |   |   |   |   |   |   altG <= 7: ACCEPT (4.79/0.37)
|   |   |   |   |   |   |   |   |   |   altG > 7
|   |   |   |   |   |   |   |   |   |   |   t_del_countG <= 4
|   |   |   |   |   |   |   |   |   |   |   |   SB2G <= 15: DISCARD (15.22/0.01)
|   |   |   |   |   |   |   |   |   |   |   SB2G > 15
|   |   |   |   |   |   |   |   |   |   |   |   map_Q0_readsG <= 0: ACCEPT
(3.88/0.88)
|   |   |   |   |   |   |   |   |   |   |   |   map_Q0_readsG > 0: DISCARD
(4.43/1.0)
|   |   |   |   |   |   |   |   |   |   |   t_del_countG > 4: ACCEPT (4.26/1.26)
|   |   |   |   |   |   |   |   |   dbsnp_siteG = COSMIC: DISCARD (0.0)
|   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP+COSMIC: DISCARD (2.44/0.0)
|   |   |   |   |   |   |   t_ref_countX > 23
```

```
|   |   |   |   |   |   |   |   n_ref_countG <= 34: DISCARD (139.34/22.09)
|   |   |   |   |   |   |   |   n_ref_countG > 34: ACCEPT (7.16/1.5)
|   |   |   |   |   |   map_Q0_readsG > 3
|   |   |   |   |   |   |   tot_depthG <= 21: DISCARD (14.92/2.29)
|   |   |   |   |   |   |   tot_depthG > 21
|   |   |   |   |   |   |   |   powerX <= 0.999968: ACCEPT (51.5/3.81)
|   |   |   |   |   |   |   |   powerX > 0.999968
|   |   |   |   |   |   |   |   |   snp_qualG <= 15.012: DISCARD (6.81/0.17)
|   |   |   |   |   |   |   |   |   snp_qualG > 15.012
|   |   |   |   |   |   |   |   |   |   normal_fG <= 0.204082: ACCEPT (18.62/3.95)
|   |   |   |   |   |   |   |   |   |   normal_fG > 0.204082: DISCARD (4.06/1.0)
|   |   |   |   |   init_n_lodG > -5.406265
|   |   |   |   |   |   t_ref_max_mapqG <= 41
|   |   |   |   |   |   |   normal_power_nspG <= 0.992708: DISCARD (17.05/0.89)
|   |   |   |   |   |   |   normal_power_nspG > 0.992708: ACCEPT (10.4/4.19)
|   |   |   |   |   |   t_ref_max_mapqG > 41
|   |   |   |   |   |   |   n_q20_countG <= 40
|   |   |   |   |   |   |   |   normal_fX <= 0.226087
|   |   |   |   |   |   |   |   |   init_t_lodX <= -5.057324
|   |   |   |   |   |   |   |   |   |   t_ins_countX <= 2
|   |   |   |   |   |   |   |   |   |   |   normal_powerX <= 0.999806: DISCARD
(14.11/4.94)
|   |   |   |   |   |   |   |   |   |   |   normal_powerX > 0.999806: ACCEPT
(99.24/22.01)
|   |   |   |   |   |   |   |   |   |   t_ins_countX > 2: DISCARD (5.79/0.38)
|   |   |   |   |   |   |   |   |   init_t_lodX > -5.057324: ACCEPT (366.26/34.86)
|   |   |   |   |   |   |   |   normal_fX > 0.226087
|   |   |   |   |   |   |   |   |   coveredX = COVERED
|   |   |   |   |   |   |   |   |   |   coveredG = COVERED
|   |   |   |   |   |   |   |   |   |   |   snp_filterG <= 0: ACCEPT (37.14/15.56)
|   |   |   |   |   |   |   |   |   |   |   snp_filterG > 0: DISCARD (2.28/0.03)
|   |   |   |   |   |   |   |   |   |   coveredG = UNCOVERED
|   |   |   |   |   |   |   |   |   |   |   t_alt_max_mapqX <= 43: ACCEPT (5.28/0.68)
|   |   |   |   |   |   |   |   |   |   |   t_alt_max_mapqX > 43: DISCARD (12.37/0.84)
|   |   |   |   |   |   |   |   |   coveredX = UNCOVERED: ACCEPT (16.57/4.6)
|   |   |   |   |   |   |   n_q20_countG > 40
|   |   |   |   |   |   |   |   tumor_fG <= 0.248366
|   |   |   |   |   |   |   |   |   normal_best_gtG = AA
|   |   |   |   |   |   |   |   |   |   init_n_lodG <= 13.105931: ACCEPT (11.0/1.12)
|   |   |   |   |   |   |   |   |   |   init_n_lodG > 13.105931: DISCARD (3.48/0.0)
|   |   |   |   |   |   |   |   |   normal_best_gtG = GG
|   |   |   |   |   |   |   |   |   |   snp_qualG <= 17.671: DISCARD (4.72/0.14)
|   |   |   |   |   |   |   |   |   |   snp_qualG > 17.671: ACCEPT (10.66/2.35)
|   |   |   |   |   |   |   |   |   normal_best_gtG = CC
|   |   |   |   |   |   |   |   |   |   SB4G <= 0: DISCARD (9.55/0.01)
|   |   |   |   |   |   |   |   |   |   SB4G > 0
|   |   |   |   |   |   |   |   |   |   |   t_lod_fstar_forG <= 5.974369: ACCEPT
(4.05/0.04)
|   |   |   |   |   |   |   |   |   |   |   t_lod_fstar_forG > 5.974369: DISCARD
(2.02/0.0)
|   |   |   |   |   |   |   |   |   normal_best_gtG = TT
|   |   |   |   |   |   |   |   |   |   dbsnp_siteG = NOVEL: DISCARD (2.79/0.0)
|   |   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP
|   |   |   |   |   |   |   |   |   |   |   t_alt_max_mapqG <= 45: ACCEPT (5.5/0.06)
|   |   |   |   |   |   |   |   |   |   |   t_alt_max_mapqG > 45: DISCARD (2.02/0.0)
|   |   |   |   |   |   |   |   |   |   dbsnp_siteG = COSMIC: ACCEPT (0.0)
|   |   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP+COSMIC: ACCEPT (1.01/0.01)
|   |   |   |   |   |   |   |   |   normal_best_gtG = AG
|   |   |   |   |   |   |   |   |   |   normal_fG <= 0.065934: ACCEPT (2.02/0.02)
|   |   |   |   |   |   |   |   |   |   normal_fG > 0.065934: DISCARD (4.94/0.0)
|   |   |   |   |   |   |   |   |   normal_best_gtG = GT: DISCARD (0.0)
|   |   |   |   |   |   |   |   |   normal_best_gtG = CT: ACCEPT (3.48/0.48)
|   |   |   |   |   |   |   |   |   normal_best_gtG = AT: DISCARD (1.46/0.0)
|   |   |   |   |   |   |   |   |   normal_best_gtG = AC: DISCARD (4.05/0.0)
```

```
|   |   |   |   |   |   |   |   |   normal_best_gtG = CG: DISCARD (2.02/0.0)
|   |   |   |   |   |   |   |   tumor_fG > 0.248366: ACCEPT (15.05/0.6)
|   |   |   normal_fG > 0.218014
|   |   |   |   normal_fX <= 0.28
|   |   |   |   |   t_q20_countX <= 17
|   |   |   |   |   |   tumor_powerG <= 0.58293: DISCARD (12.45/1.05)
|   |   |   |   |   |   tumor_powerG > 0.58293: ACCEPT (22.67/4.49)
|   |   |   |   |   t_q20_countX > 17
|   |   |   |   |   |   n_ref_countG <= 16: DISCARD (209.28/18.62)
|   |   |   |   |   |   n_ref_countG > 16
|   |   |   |   |   |   |   normal_power_nspX <= 0.999931: ACCEPT (5.82/0.76)
|   |   |   |   |   |   |   normal_power_nspX > 0.999931: DISCARD (35.35/6.33)
|   |   |   |   normal_fX > 0.28: DISCARD (465.49/7.26)
|   |   total_readsG > 1434
|   |   |   SB4G <= 5: DISCARD (1054.32/5.69)
|   |   |   SB4G > 5
|   |   |   |   AT_FmadG <= 7.5: DISCARD (22.24/0.01)
|   |   |   |   AT_FmadG > 7.5: ACCEPT (4.04/0.04)
|   n_alt_sumG > 202
|   |   normal_fG <= 0.287356
|   |   |   total_readsG <= 1744
|   |   |   |   snp_qualG <= 74.505
|   |   |   |   |   normal_fG <= 0.214431
|   |   |   |   |   |   snp_qualG <= 12.162
|   |   |   |   |   |   |   coveredX = COVERED: DISCARD (284.27/3.85)
|   |   |   |   |   |   |   coveredX = UNCOVERED
|   |   |   |   |   |   |   |   SB3G <= 86: DISCARD (12.41/1.5)
|   |   |   |   |   |   |   |   SB3G > 86: ACCEPT (3.09/0.08)
|   |   |   |   |   |   snp_qualG > 12.162
|   |   |   |   |   |   |   tumor_fG <= 0.231579
|   |   |   |   |   |   |   |   SB2X <= 97
|   |   |   |   |   |   |   |   |   judgementX = KEEP: ACCEPT (3.48/0.31)
|   |   |   |   |   |   |   |   |   judgementX = REJECT
|   |   |   |   |   |   |   |   |   |   normal_fX <= 0.26
|   |   |   |   |   |   |   |   |   |   |   map_Q0_readsG <= 118
|   |   |   |   |   |   |   |   |   |   |   |   t_alt_countG <= 16
|   |   |   |   |   |   |   |   |   |   |   |   |   snp_qualG <= 36.873
|   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtX = TT: DISCARD
(4.74/0.06)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtX = AA: DISCARD
(2.67/0.02)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtX = GG
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   t_lod_fstar_revX <=
4.93973: DISCARD (9.0/0.09)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   t_lod_fstar_revX >
4.93973: ACCEPT (2.21/0.18)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtX = CC: DISCARD
(10.15/1.1)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtX = CT
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   n_q20_countG <= 42:
DISCARD (22.77/1.17)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   n_q20_countG > 42
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   init_t_lodX <=
-4.544543
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   total_readsG <=
119: ACCEPT (2.5/0.07)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   total_readsG >
119: DISCARD (11.84/0.13)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   init_t_lodX >
-4.544543: ACCEPT (22.8/8.17)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtX = AG
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   SB3G <= 3
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tumor_fX <= 0.126984
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   coveredX = COVERED
```

```
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   altG <= 11:
DISCARD (9.32)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   altG > 11
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   snp_qualX
<= 30.941: DISCARD (10.47/1.78)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   snp_qualX
> 30.941: ACCEPT (5.55/1.14)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   coveredX =
UNCOVERED: ACCEPT (4.35/1.32)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tumor_fX > 0.126984:
ACCEPT (4.43/0.35)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   SB3G > 3: DISCARD
(11.0/0.15)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtX = CG
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tumor_powerG <= 0.9982:
DISCARD (10.9/0.07)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tumor_powerG > 0.9982
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteX = NOVEL:
ACCEPT (2.21/0.18)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteX = DBSNP
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   init_n_lodG <=
-14.65134: DISCARD (6.72/1.11)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   init_n_lodG >
-14.65134: ACCEPT (3.2/0.2)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteX = COSMIC:
ACCEPT (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteX = DBSNP
+COSMIC: ACCEPT (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtX = AT
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   contaminant_lodG <=
-0.420604: ACCEPT (2.01/0.01)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   contaminant_lodG >
-0.420604: DISCARD (8.95/0.1)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtX = GT
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   map_Q0_readsX <= 7:
DISCARD (7.67/2.07)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   map_Q0_readsX > 7: ACCEPT
(2.67/0.21)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtX = AC
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_power_wspG <=
0.999912: ACCEPT (2.3/0.25)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_power_wspG >
0.999912: DISCARD (16.6/3.11)
|   |   |   |   |   |   |   |   |   |   |   |   |   snp_qualG > 36.873
|   |   |   |   |   |   |   |   |   |   |   |   |   |   coveredX = COVERED
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteG = NOVEL
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtG = AA:
ACCEPT (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtG = GG:
ACCEPT (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtG = CC:
ACCEPT (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtG = TT:
ACCEPT (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtG = AG
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   t_lod_fstar_revG
<= 3.614914: ACCEPT (4.68/0.2)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   t_lod_fstar_revG >
3.614914: DISCARD (2.18/0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtG = GT:
DISCARD (3.07/0.15)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtG = CT
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tot_depthG <= 142:
ACCEPT (5.49/1.29)
```

```
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tot_depthG > 142:
DISCARD (3.3/0.12)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtG = AT:
ACCEPT (5.0/1.52)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtG = AC:
ACCEPT (1.17/0.17)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   normal_best_gtG = CG:
ACCEPT (1.49/0.49)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   t_ref_max_mapqG <= 41:
DISCARD (2.02/0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   t_ref_max_mapqG > 41:
ACCEPT (29.11/6.63)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteG = COSMIC:
ACCEPT (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP
+COSMIC: DISCARD (1.01/0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   coveredX = UNCOVERED
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   p_allX <= -1.199: DISCARD
(8.33/0.21)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   p_allX > -1.199: ACCEPT
(4.18/1.07)
|   |   |   |   |   |   |   |   |   |   |   |   |   t_alt_countG > 16
|   |   |   |   |   |   |   |   |   |   |   |   |   |   SB4X <= 4: DISCARD (38.72/1.75)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   SB4X > 4
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   n_alt_countG <= 11: ACCEPT
(5.01/1.01)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   n_alt_countG > 11
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   SB2X <= 28: ACCEPT
(4.08/1.4)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   SB2X > 28: DISCARD
(14.76/0.34)
|   |   |   |   |   |   |   |   |   |   |   |   map_Q0_readsG > 118: DISCARD (30.89/1.18)
|   |   |   |   |   |   |   |   |   |   |   normal_fX > 0.26
|   |   |   |   |   |   |   |   |   |   |   |   map_Q0_readsX <= 3: DISCARD (55.71/0.7)
|   |   |   |   |   |   |   |   |   |   |   |   map_Q0_readsX > 3
|   |   |   |   |   |   |   |   |   |   |   |   |   t_del_countX <= 22
|   |   |   |   |   |   |   |   |   |   |   |   |   |   t_alt_max_mapqG <= 22
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   t_ref_max_mapqG <= 55: DISCARD
(3.12)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   t_ref_max_mapqG > 55
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteG = NOVEL:
DISCARD (2.57/0.14)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP:
ACCEPT (4.87/0.82)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteG = COSMIC:
ACCEPT (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dbsnp_siteG = DBSNP
+COSMIC: ACCEPT (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   t_alt_max_mapqG > 22: DISCARD
(18.85/0.19)
|   |   |   |   |   |   |   |   |   |   |   |   t_del_countX > 22: ACCEPT (2.22/0.2)
|   |   |   |   |   |   |   |   SB2X > 97: DISCARD (181.93/10.29)
|   |   |   |   |   |   |   tumor_fG > 0.231579
|   |   |   |   |   |   |   |   normal_fX <= 0.184783
|   |   |   |   |   |   |   |   |   n_alt_countG <= 12: ACCEPT (34.47/5.51)
|   |   |   |   |   |   |   |   |   n_alt_countG > 12
|   |   |   |   |   |   |   |   |   |   Power_Pos_SBG <= 0.999993: DISCARD
(10.55/1.01)
|   |   |   |   |   |   |   |   |   |   Power_Pos_SBG > 0.999993
|   |   |   |   |   |   |   |   |   |   |   AT_FmadG <= 11.5: DISCARD (5.57/0.01)
|   |   |   |   |   |   |   |   |   |   |   AT_FmadG > 11.5: ACCEPT (7.08/1.07)
|   |   |   |   |   |   |   |   normal_fX > 0.184783
|   |   |   |   |   |   |   |   |   tumor_fX <= 0.303571: DISCARD (22.65/1.09)
```

```
|   |   |   |   |   |   |   |   |   tumor_fX > 0.303571
|   |   |   |   |   |   |   |   |   |   normal_best_gtX = TT: DISCARD (0.0)
|   |   |   |   |   |   |   |   |   |   normal_best_gtX = AA: DISCARD (0.0/0.0)
|   |   |   |   |   |   |   |   |   |   normal_best_gtX = GG: DISCARD (0.02/0.0)
|   |   |   |   |   |   |   |   |   |   normal_best_gtX = CC: DISCARD (0.03/0.0)
|   |   |   |   |   |   |   |   |   |   normal_best_gtX = CT: DISCARD (7.98/1.38)
|   |   |   |   |   |   |   |   |   |   normal_best_gtX = AG
|   |   |   |   |   |   |   |   |   |   |   n_ref_countG <= 55: ACCEPT (4.45/0.02)
|   |   |   |   |   |   |   |   |   |   |   n_ref_countG > 55: DISCARD (4.47/0.0)
|   |   |   |   |   |   |   |   |   |   normal_best_gtX = CG: ACCEPT (1.11/0.06)
|   |   |   |   |   |   |   |   |   |   normal_best_gtX = AT: DISCARD (1.28/0.06)
|   |   |   |   |   |   |   |   |   |   normal_best_gtX = GT: DISCARD (3.51/0.75)
|   |   |   |   |   |   |   |   |   |   normal_best_gtX = AC: ACCEPT (1.12/0.06)
|   |   |   |   |   |   |   |   |   normal_fG > 0.214431: DISCARD (795.6/36.35)
|   |   |   |   |   |   |   |   snp_qualG > 74.505
|   |   |   |   |   |   |   |   |   normal_fG <= 0.205732
|   |   |   |   |   |   |   |   |   |   t_del_countG <= 72
|   |   |   |   |   |   |   |   |   |   |   tumor_powerG <= 0.999407
|   |   |   |   |   |   |   |   |   |   |   |   normal_fX <= 0.137931: ACCEPT (4.3/0.91)
|   |   |   |   |   |   |   |   |   |   |   |   normal_fX > 0.137931: DISCARD (8.27/0.19)
|   |   |   |   |   |   |   |   |   |   |   tumor_powerG > 0.999407: ACCEPT (223.66/29.41)
|   |   |   |   |   |   |   |   |   |   t_del_countG > 72
|   |   |   |   |   |   |   |   |   |   |   n_alt_countG <= 7: ACCEPT (2.02/0.02)
|   |   |   |   |   |   |   |   |   |   |   n_alt_countG > 7: DISCARD (13.29/0.01)
|   |   |   |   |   |   |   |   |   normal_fG > 0.205732
|   |   |   |   |   |   |   |   |   |   n_alt_sumX <= 97: ACCEPT (9.91/1.72)
|   |   |   |   |   |   |   |   |   |   n_alt_sumX > 97
|   |   |   |   |   |   |   |   |   |   |   t_ref_max_mapqG <= 57
|   |   |   |   |   |   |   |   |   |   |   |   normal_fX <= 0.260465: ACCEPT (3.04/0.04)
|   |   |   |   |   |   |   |   |   |   |   |   normal_fX > 0.260465: DISCARD (5.79)
|   |   |   |   |   |   |   |   |   |   |   t_ref_max_mapqG > 57: DISCARD (156.06/14.09)
|   |   |   |   total_readsG > 1744
|   |   |   |   |   t_alt_countG <= 152: DISCARD (3569.33/35.33)
|   |   |   |   |   t_alt_countG > 152
|   |   |   |   |   |   n_ref_sumG <= 29056: DISCARD (47.51/1.03)
|   |   |   |   |   |   n_ref_sumG > 29056: ACCEPT (6.07/1.06)
|   |   |   normal_fG > 0.287356
|   |   |   |   n_alt_sumX <= 67
|   |   |   |   |   judgementX = KEEP: ACCEPT (7.05/2.32)
|   |   |   |   |   judgementX = REJECT
|   |   |   |   |   |   map_Q0_readsX <= 1
|   |   |   |   |   |   |   tumor_powerX <= 0.292001: ACCEPT (16.74/5.62)
|   |   |   |   |   |   |   tumor_powerX > 0.292001: DISCARD (36.76/4.26)
|   |   |   |   |   |   map_Q0_readsX > 1: DISCARD (67.27/2.04)
|   |   |   |   n_alt_sumX > 67: DISCARD (17906.84/66.65)
```

Chapter Four

Identifying Short Tandem Repeat Genetic Markers for Lung Squamous Cell Carcinoma

4.1 Abstract

Short tandem repeats are repetitive units of 2-6 nucleotides found interspersed in the genome and have been traditionally used as genetic markers in population genetics. To investigate their contribution to cancer, we analyze 285,063 short tandem repeat regions from the exomes of 32 matched normal and tumor lung tissues obtained from The Cancer Genome Atlas. We identify 103 STRs that vary between cancer and normal tissues but not between normal and 1000 Genomes data. We study the possible functional impact of these variations by analyzing the mRNA gene expression data.

Keywords: Short tandem repeat variation, lung cancer, microsatellite instability

4.2 Introduction

Lung cancer is the leading cause of cancer-related deaths in males and females [104, 105]. Lung cancer can be classified into two main types: small cell lung cancer (SCLC) and non-small cell lung carcinoma (NSCLC), which account for 20% and 80% of lung cancer, respectively [106]. Cancer cells in SCLC are small, highly aggressive, and rapidly metastasize to other parts of the body [107]. The median survival time of a person diagnosed with SCLC is only 12-16 months [108]. NSCLC can be further classified into three histological subtypes: adenocarcinoma, lung squamous cell carcinoma (LUSC), and large cell lung cancer. LUSC affects the central part of the lungs, near the upper airways [109], accounting for about 30% of lung cancer [110] (i.e., 37% of NSCLC). It has been found that LUSC is highly associated with people who have smoking history and occurs more frequently in males than in females [109].

Primary tumors can also be classified into synchronous lung cancer and metachronous lung cancer. Tumors of synchronous lung cancer develop at multiple sites simultaneously, whereas tumors of metachronous lung cancer develop progressively, mostly through metastasis. Distinguishing these two types is important from a prognostic and therapeutic perspective. In a clinical setting, multiple tumors that are isolated but histologically similar occur in only 0.2 - 2.0% of the patients in the case of lung cancer [111].

Short tandem repeats (STRs) are short segments of DNA of lengths 2-6 nucleotides that repeat a varying number of times throughout the genome. STRs are abundant throughout the genome and are known for their high polymorphism. The mutation rate of STRs is influenced by

the purity (i.e., how perfectly the STR gets repeated due to mutations) and length of repeat sequence [112]. The mutation rate of STRs is 1.5×10^{-2} per locus per gamete per generation, which is 200-times higher than that of copy number variations (CNV) [113] and 200,000 times higher than that of single nucleotide polymorphisms (SNPs) [114]. This genomic instability makes STRs ideal for genetic fingerprinting, genotyping, paternity testing, and forensic analysis. STR variations have been found associated with over 40 human hereditary diseases such as Huntington's disease, fragile X syndrome, myotonic muscular dystrophy, and neurological diseases [115].

Analyzing STRs and identifying STR variations through Next-Generation Sequencing (NGS) is challenging due to difficulties in sequencing, mapping, and studying the variants using standard techniques. For example, PCR amplification of an STR region could introduce replication errors due to DNA replication slippage events despite the use of a high fidelity DNA polymerase [116]. While mapping reads to regions that contain indels and repeats, there is an increase in run time due to gapped alignments [69]. Genotype calling in STR region becomes problematic too. Reads must encompass an entire STR region and its corresponding flanking regions to confidently support a genotype [112], since calling an STR genotype from reads that do not fully encompass an STR region could lead to identifying a wrong genotype. Consequently, STR mutations cannot be called accurately by variant callers for indels and requires additional investigations. There have been several tools developed for STR calling in the whole genome sequencing data. LobSTR calls STRs using three steps, i.e., (i) sensing - identify the motif that repeats in the locus (ii) alignment- realign the reads in the location where the STR

is present, and (iii) allelotyping - determine the number of times the motif repeats thereby, calling the STR allele [69]. RepeatSeq uses a Bayesian model approach guided by an error model to determine the STR allele [70]. The STR alleles called by both Repeatseq and LobSTR are hosted on 1000 Genomes website [23]. It has previously been shown that LobSTR works better than RepeatSeq in determining STR alleles [71].

STRs are also called microsatellites. Microsatellite instability (MSI) is a known term that has been associated with various cancers previously [117]. MSI is caused by hypermutability of microsatellite regions due to slippage of DNA polymerase and impaired function of mismatch repair protein. DNA mismatch repair gene deficit causing MSI is a known occurrence in the prognosis and development of colorectal cancer [118]. MSI testing of four mismatch repair genes (MLH1, MSH2, PMS2 and MSH6) is performed on patients with colon cancer for prognosis of hereditary nonpolyposis colorectal cancer [119]. MSI has also been associated with lung cancer [111]. Approximately 45% of small cell lung cancer is associated with MSI, while 2-72% of NSCLC is associated with MSI [120] . The hypermutability of microsatellites compared to regions that do not have repetitive DNA sequences can be used as a prognostic tool for different cancers. No one has yet characterized all the STR regions that vary in a particular cancer [121].

In this work, we examine the STR regions that vary in lung cancer and try to identify those that could have a functional impact. We compare tumor and normal exomes of 32 individuals with LUSC. The tumor and normal exomes are obtained from the same primary

tissue. These matched tissue samples provide sufficient information for studying the functional impact of STR variation on LUSC. Then, we check whether STR variations are present within a gene that exhibits differential expression across matched samples (Figure 4.1).

4.3 Results

The study sample consists of 23 males and 9 females from the Caucasian population. The clinical information of these 32 individuals is shown in Supplementary Table S4.1. A total of 285,063 STR regions of normal and tumor exomes of 32 individuals were genotyped using LobSTR [69]. Across individuals, we observed different completeness (proportion of called regions). The average number of called regions across the 32 individuals is 23,106 (SD±8,137) in normal samples and 24,456 (SD±16,556) in tumor samples. Figure 4.2 summarizes the number of homozygous and heterozygous regions in normal and tumor samples. On average, there are 21,903 (SD±7795) homozygous and 1,203 (SD±402) heterozygous regions called in normal samples, and 23,285 (SD±16035) homozygous and 1,171 (SD±538) heterozygous regions called in tumor samples of the 32 individuals.

The numbers of concordant and discordant (including completely discordant and partially discordant, see methods for definition) sites in each individual are shown in Table 4.1. Concordant sites indicate STR commonality between the tumor and normal genotypes, whereas completely discordant sites and partially discordant sites are STRs with different number of repeats in tumor and normal genotypes and haplotypes, respectively. As shown in Table 4.1, an average of 1,080 sites, accounting for 10% (SD±1.2%) of the total sites called in each individual,

are discordant. Out of these discordant sites, an average of 575 sites are completely discordant, accounting for 26% (SD±10%) of the total number of discordant sites called in each individual.

4.3.1 Analysis of STR regions

To find STR regions that might be linked to LUSC, we conducted three statistical tests of copy number differences between: (i) tumor versus normal samples, (ii) tumor samples versus 1000 Genomes data, and (iii) normal samples versus 1000 Genomes data. Only the European population in the 1000 Genomes project was considered in the analyses to be consistent with the race of the 32 individuals with LUSC (see Methods section).

In the analysis of tumor versus normal, a paired Wilcoxon test found 1,087 STR regions with p-value less than 0.1. In the other two analyses, an unpaired Wilcoxon test (see Methods for details) found 120,158 regions with p-value lesser than 0.1 when comparing normal with 1000 Genomes data, and 130,464 regions with p-value lesser than 0.1 when comparing tumor with 1000 Genomes samples. Figure 4.3 shows a distribution of the p-values of total 285,063 STR regions with the three statistical analyses.

STR regions can potentially be considered as tumor markers if STR repetition numbers are significantly different for the tumor versus normal and the tumor versus 1000 Genomes comparison but not for the normal versus 1000 Genomes comparison (Table 4.2). Setting a p-value threshold of 0.1 found 103 such STR regions. Out of the 285,063 STR regions that were analyzed, 143,473 STRs were present within genes. Out of the 103 candidate cancer marker STR

regions, 58 were present within genes. Table 4.3 lists the 58 STRs that reside in 58 genes (the remaining 45 STRs are shown in Supplementary Table S4.2).

Another interesting category includes the STRs with repetition numbers significantly different for the tumor versus the 1000 Genomes comparison, and normal versus 1000 Genomes comparison but not for the tumor versus normal comparison (Table 4.2). These STR regions could be considered as potential genetic markers for cancer propensity or predisposition. Our analysis identified 12,048 such STR sites, 8,301 of which are within 6,057 genes. To make sure that these genes could be considered as candidates for propensity for lung cancer, we checked for genes that might be associated with cancer. The 6,057 genes were compared with DisGeNet [29], a database that associates genes with diseases. There were 197 genes that were associated with any cancer in DisGeNet. There were 296 STRs present within these 197 genes. We also did a functional annotation to find out whether the genes were associated to any pathways related to any cancer (See Functional Annotation Analysis). Thus, out of the 6,057 genes that contained STR variation between tumor versus 1000G and normal versus 1000G, there were 197 genes associated with any cancer in DisGeNet (Supplementary Table S4.3).

4.3.2 Gene Expression Analysis

The 32 samples in this study were selected on the basis that the individuals have both whole exome and RNASeq data sequenced. The tool EBSeq was used to identify genes that were significantly differentially expressed (i.e., with a fold change of greater than 1.5 and an FDR

value lesser than 0.05) between normal and tumor samples (details in Methods). Out of the total 20,501 genes that were analyzed, 3,797 are significantly differentially expressed. Comparing the 58 genes containing tumor causing STR regions (Table 4.3) with the 3,797 genes that were significantly differentially expressed, we found 51 genes in intersection, which cover 25 STR regions, and there is a significant enrichment of genes that have differential gene expression in the potential STR markers for cancer (hypergeometric test, p-value = 1.29e-07).

4.3.3 Functional Annotation Analysis

A functional annotation was done on all the 58 genes using ConsensusPathDB [122]. The enriched pathways include extracellular matrix (ECM) receptor interaction, ECM organization, miRNA targets in ECM and membrane receptors, cell adhesion molecules (CAMs), neural cell adhesion molecule (NCAM1) interactions, signaling events regulated by Ret tyrosine kinase, binding and uptake of ligands by scavenger receptors, spliceosome, Syndecan-1-mediated signaling events, and Vitamin B12 metabolism. (Supplementary Table S4.4). Thirty three genes from the input list of 58 genes were present in at least one of the above mentioned pathways. The p-value cutoff for these pathways is 0.01.

We performed GO analysis and found the terms associated with the gene list, including learned vocalization behavior or vocal learning neuron differentiation and development, neuron projection guidance and development, axon guidance and development, auditory and vocalization behavior, cell adhesion, differentiation, morphogenesis and development, taxis, proteoglycan binding, growth-factor binding, 4 iron, 4 sulfur (Fe-S) cluster binding,

exopeptidase activity, and complex of collagen trimers and basement membrane. A list of all the GO terms with p-value less than 0.01 is given in Supplementary Table S4.5. Fifty four genes from the input list were associated with at least one of the GO terms.

A functional annotation was also carried out for the 6,057 genes indicating propensity for lung cancer. Out of the 6057 genes, 202 were associated with pathway terms related to cancer such as ‘Pathways in cancer’, 'Proteoglycans in cancer', and ‘Small cell lung cancer’. Not surprisingly, the pathways or GO terms include related cell development, cell signaling, cell-cell communication, cell transportation, cell apoptosis, regulation of immune response, protein and lipid binding required for cell cycles. There were 340 STRs within these 202 genes. Supplementary Table S4.3 shows a list of 624 STRs present within 391 genes that might have a predisposition to lung cancer (inclusive of genes from the DisGeNet analysis).

4.4 Discussion

STR is more variable in the genome compared to SNPs or indels [113], and studying STR instability in cancer is of great importance. Detection of STR instability could be used to early diagnose cancer and reduce the mortality rate. This is especially important for lung cancer, since lung cancer has the highest rate of mortality among all cancers [106]. LUSC, as an aggressive form of lung cancer, is even harder to diagnose due to its quick metastases. STR instability in LUSC could thus be used as an early detection tool. With this motivation, we present a list of 103 STR regions that are variable in LUSC compared to the healthy tissue. Out of the 103 STR regions, 58 of them are present within a gene. Twelve of the fifty eight genes have a differential

gene expression in the corresponding RNASeq tissue. We also present a list of 391 genes that might have an STR variation and could possibly have a predisposition for LUSC. Additional biological evaluation with more samples is necessary in this aspect.

Supplementary Table S4.4 lists the pathway terms that are associated with the 58 genes that have STRs with variation in LUSC tumor samples. Several pathways terms related to extracellular matrix (ECM) organization, regulation and miRNA in extracellular matrix are affected by the genes with microsatellite instability in LUSC. Tissues are composed of cells and extracellular matrix. This ECM is mainly made up of a complex mixture of macromolecules which includes polysaccharides and proteins. ECM has important roles in cell adhesion, differentiation, migration, homeostasis, and apoptosis [123]. The genes that affect ECM organization and regulation are genes that encode for collagen type IV (COL4A1), collagen type V (COL5A1), Tenascin N (TNN), and neurexin (NRXN1). TNN and NRXN1 affect degradation of extracellular matrix. TNN also affects microRNAs in cancer. Collagens are one of the most abundant proteins that make up ECM. CNTNAP1 and CNTNAP2 are members of the neurexin family and encode for proteins that affect pathways that regulate cell adhesion molecules [124]. Cell adhesion molecules are present on the surface of cells and initiate adhesion of cells to nearby cells or ECM. Syndecans are proteoglycans that play an important role in cell adhesion, proliferation, and migration. Syndecan-1 is the major syndecan of epithelial cells and has been shown to be associated with myeloma, medulloblastoma, endometrial cancer and breast cancer [125]. NCAM1 impacts cell-cell and cell-matrix interactions. Differential expression of NCAM1 affects different kinds of tumors including colon carcinoma, pancreatic cancer, astrocytoma,

neuroblastoma, and certain neuroendocrine tumors and recently with NSCLC in particular [126]. Scavenger receptors are a huge range of receptors that bind with other co-receptors to carry out various functions such as pathogen clearance, removal of modified lipoproteins, and lipid transport, among others [127]. Scavenger receptors are primarily expressed in macrophages. The class A scavenger receptor (SR-A) gene bind to apoptotic cells thus, partaking in apoptosis and inflammation. Scavenger receptors, specifically SR-A have been associated with lung cancer [128]. Oncogenic fusion of RET and tyrosine kinase receptor genes has been associated with thyroid cancer, breast cancer, and NSCLC [129]. The spliceosome is a complex of snRNAs and proteins that is involved in splicing, i.e., removal of introns from pre-mRNA. The spliceosome has been shown to play an important part in tumors due to its role in alternate splicing and can thus be considered as a prospective therapeutic target [130]. Vitamin B-12 deficiency has been shown to be associated with genome-wide DNA hypomethylation in prostate cancer and LUSC. Overexpression of vitamin B12 receptors is associated with ovary, kidney, uterus, testis, brain, colon, lung, and myelocytic blood cells cancer [131].

Supplementary Table S4.5 lists the GO terms associated with the 58 genes that have STRs with variation in LUSC tumor samples when compared to normal and 1000 Genomes population data. Proteoglycans are glycosylated proteins. Many types of proteoglycans are present in ECM. Many types of proteins, such as proteins in ECM, growth factors and cytokines, plasma proteins, transmembrane proteins and cytoplasmic proteins, bind to proteoglycans [132]. Proteoglycans have multiple functions, such as structural functions, and enzymatic activities, could act as cell surface receptors, or could have a role in tissue development and repair. Their

role in cancer has been investigated in multiple studies previously [133]. The GO term chemotaxis being associated with any cancer makes sense intuitively, as chemotaxis means the movement of cells in response to an increasing or decreasing extracellular chemical gradient. Chemotaxis in a healthy human being is required for recruitment of inflammatory cells, and organ development during embryogenesis. Chemotaxis in the case of cancer could help the proliferation and metastasis of tumor cells [134]. The aggressive nature of SCLC makes it a neuroendocrine subtype of lung cancer. SCLC cell lines have also been associated with expression of multi-lineage stem cell markers and thus, differentiating into neurons, adipocytes, and osteocytes in vitro [135]. SCLC also has characteristics associated with neurons which is not surprising since they are derived from neuroectoderm [108]. It has also been found that different neuroectodermal tumors share neural antigens with SCLC [136]. Axon development and axon guidance were other GO terms that were present in the functional annotation analysis. Axons transmit information from neurons to other neurons, muscles and glands. Axon guidance molecules have been known to regulate cell migration and apoptosis. Axon guidance molecules also regulate neuronal migration and survival. Axon guidance molecules are differentially expressed in different kinds of tumors including lung cancer [137]. There is growing evidence of "4 iron, 4 sulfur clusters" affecting DNA replication and repair [138]. Interestingly, GO terms such as "auditory and vocalization behavior" were associated with the genes. Since SCLC is highly aggressive and it metastasizes to other parts of the body, if it metastasizes to the larynx box, it could affect the vocal cord. Previously vocal cord paralysis has been shown to be associated with lung cancer [139].

A possible development of this work is to extend from STR-based testing (done in this study using Wilcoxon's test) to gene-based testing to incorporate the correlation between multiple STRs present within each gene. Linear regression, logistic regression, Fourier transformation [140], principal-components analysis [141], and cluster analysis [142] have been used previously for associating multiple correlated SNPs with a particular phenotype. It would be interesting to analyze the interaction between multiple STRs within one gene to study how different STR variations could affect a particular gene expression in the case of cancer.

Similar analysis could be performed to study the functional impact of varying STRs in other types of lung cancer. For example, there are 58 and 48 people whose exome and mRNA have been sequenced for lung adenocarcinoma (LUAD) and lung hepatocellular carcinoma (LIHC) respectively. It would be interesting to see if the same STR regions or the same genes are affected due to microsatellite instability in different cases of lung cancer. The differences in the STR regions or genes affected should also provide some insight into the classification of subtypes of lung cancer.

4.5 Methods

4.5.1 Real Data

We obtained LUSC data from TCGA. The raw data contains 502 individuals, among which 227 have exome sampled from the primarily affected site, i.e., they have both normal and tumor samples from the lung tissue. We further filtered out individuals who have no gene expression data extracted from lung tissues and retained 44 individuals. Among them, 32 came

from the Caucasian population, 3 were African American and the race of the rest of the 9 people was unknown. To minimize the confounding effect of population stratification, we only considered the 32 Caucasians. For each STR, based on the number of times that the STR appears, that is, STR genotypes, in both normal and cancer tissues, we defined three basic classes of STRs, concordant STRs, partially concordant STRs, and discordant STRs. Concordant STRs are those STRs that are homozygous and have the same copy numbers for both tumor and normal tissues. For example, if a STR has a homozygous genotype and repeats 22 times (i.e., genotype: 22/22) in both normal and tumor samples, it is considered as a concordant site; if a STR's genotype is 22/22 in normal sample and 22/23 in tumor sample, it is considered as partially discordant; if a STR's genotype is 22/22 in normal sample and 23/23 in tumor sample, it is considered as completely discordant.

4.5.2 STR regions

LobSTR [69] was run on the normal and tumor samples to get the variability in STR regions. LobSTR was chosen over another popular repeat determining tool RepeatSeq as LobSTR was concluded to be the more superior tool [71]. Marshfield panel consists of 485 autosomal STR markers from 105 individuals from the 1000 Genomes population. Willems et. al. compared the LobSTR and RepeatSeq calls to the genotypes generated from the Marshfield panel. A correlation analysis showed that LobSTR outperformed RepeatSeq with R^2 values of 0.71 compared to 0.4 of RepeatSeq and hence, LobSTR was chosen for our further analyses. LobSTR requires the BAM file and the STR region file as input. It calculates the variability in STR by first identifying the repeat motif, locally realigning the reads in the STR region to

accurately generate a CIGAR score for the reads and then determining the number of times the motif repeats. We examined approximately 700,000 STR regions that were found previously [71]. Out of these STRs, 285,063 were present in the exon regions of the normal, tumor, and 1000 genome samples.

4.5.3 STR Analysis in Exomes

The STR regions with significant variability in the number of motif repeats in the normal and tumor samples were identified using paired Wilcoxon test at nominal level 0.1. The value 0.1 was chosen to include as many STRs as possible since the number of STRs would be further filtered by gene expression (Section 4.5.4) and functional annotation analyses (Section 4.5.5). For the analyses between (i) tumor versus 1000 Genomes, and (ii) normal versus 1000 Genomes, an unpaired Wilcoxon test was used. Since there were 345 individuals of European origin in the 1000 Genomes population, we resampled the population such that 32 STRs were randomly chosen from the European population to compare with the 32 normal or 32 tumor samples. Once resampled, the STRs from the European population of the 1000 Genomes were compared to the normal and tumor samples using unpaired Wilcoxon test.

4.5.4 Gene Expression Analysis

We used TCGA2STAT [143] to download the RNASeq version 2 data of the 32 individuals. A standard pipeline to determine the expression levels in RNASeq version 2 data is to use Mapslice [144] to do read alignment and use RSEM [145] to perform quantification. The downloaded RNASeq version 2 data was then analyzed using the EBSeq package [146]. We used

the intersectBed function in BedTools to check whether STRs are present within a gene. The gene annotation information was retrieved from UCSC (ucscHg19). All gene co-ordinates were downloaded from the UCSC genome browser. The minimum starting co-ordinate and maximum ending co-ordinate were recorded for genes with multiple entries. A total of 30,000 genes are included in the gene annotation file. This procedure yields a total of 260,885 STRs which are present within the genes.

4.5.5 Functional Annotation

A functional annotation analysis of 58 genes was done using ConsensusPathDB [122]. ConsensusPathDB-human is a database that integrates various interaction networks obtained from 32 publicly available resources. These interactions include protein interactions, signaling reactions, metabolic reactions, gene regulations, genetic interactions, drug-target interactions and biochemical pathways in Homo sapiens.

4.6 Conclusion

We provide a list of 103 STRs that vary for LUSC and could potentially be used as cancer markers. We also provide a list of 624 STRs present within 391 genes that have been linked to cancer in the past. These 624 STRs could potentially be used as predisposition markers for LUSC. These candidate STRs require further experimental validation, which is beyond the scope of our current computational analysis.

4.7 Abbreviations

CNV: copy number variations

COL4A1: Collagen type IV

COL5A1: Collagen type V

ECM: Extra-Cellular Matrix

GO: Gene Ontology

LIHC: Lung Hepatocellular Carcinoma

LUAD: Lung Adenocarcinoma

LUSC: Lung Squamous Cell Carcinoma

NRXN1: Neurexin

NSCLC: Non-Small Cell Lung Carcinoma

SCLC Small Cell Lung Carcinoma

SR-A: Scavenger Receptor (SR-A)

TCGA: The Cancer Genome Atlas

TNN: Tenascin N

SNP: Single Nucleotide Polymorphism

STR: Short Tandem Repeat

Figures

Figure 4.1: Flowchart of STR analysis procedure.

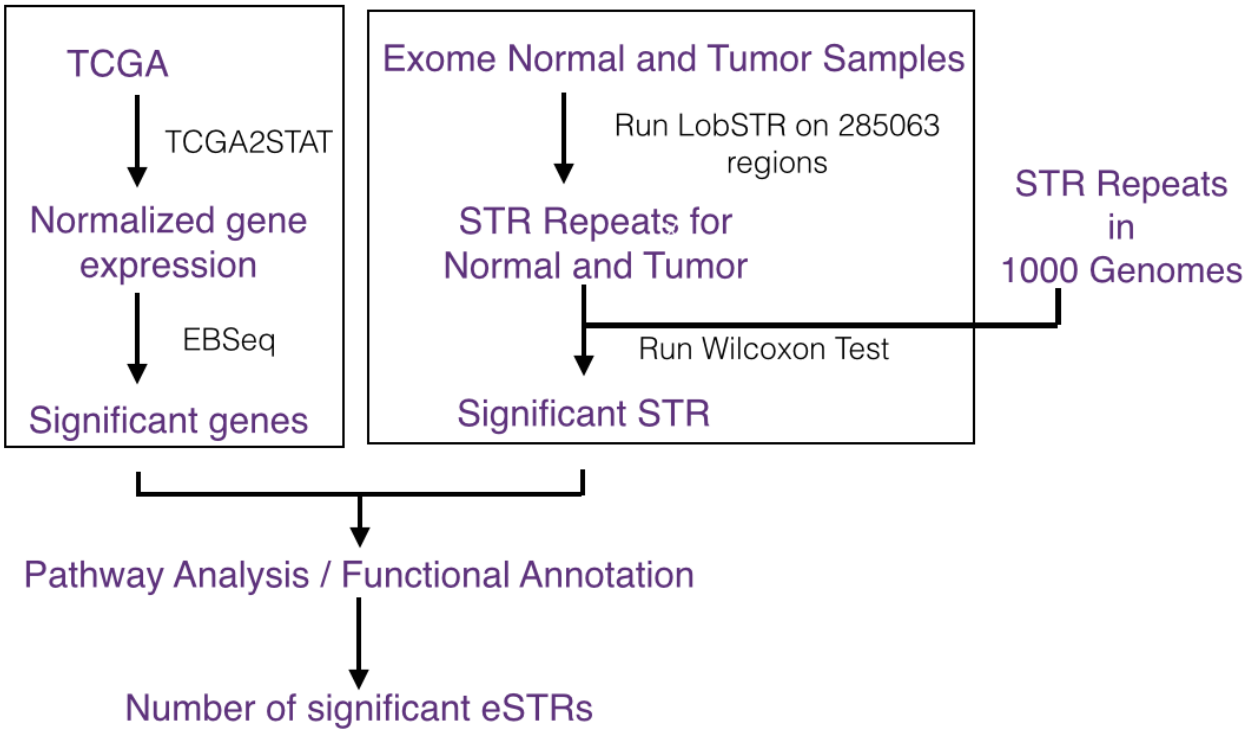


Figure 4.2: Average number of homozygous and heterozygous regions in normal and tumor samples of the 32 individuals.

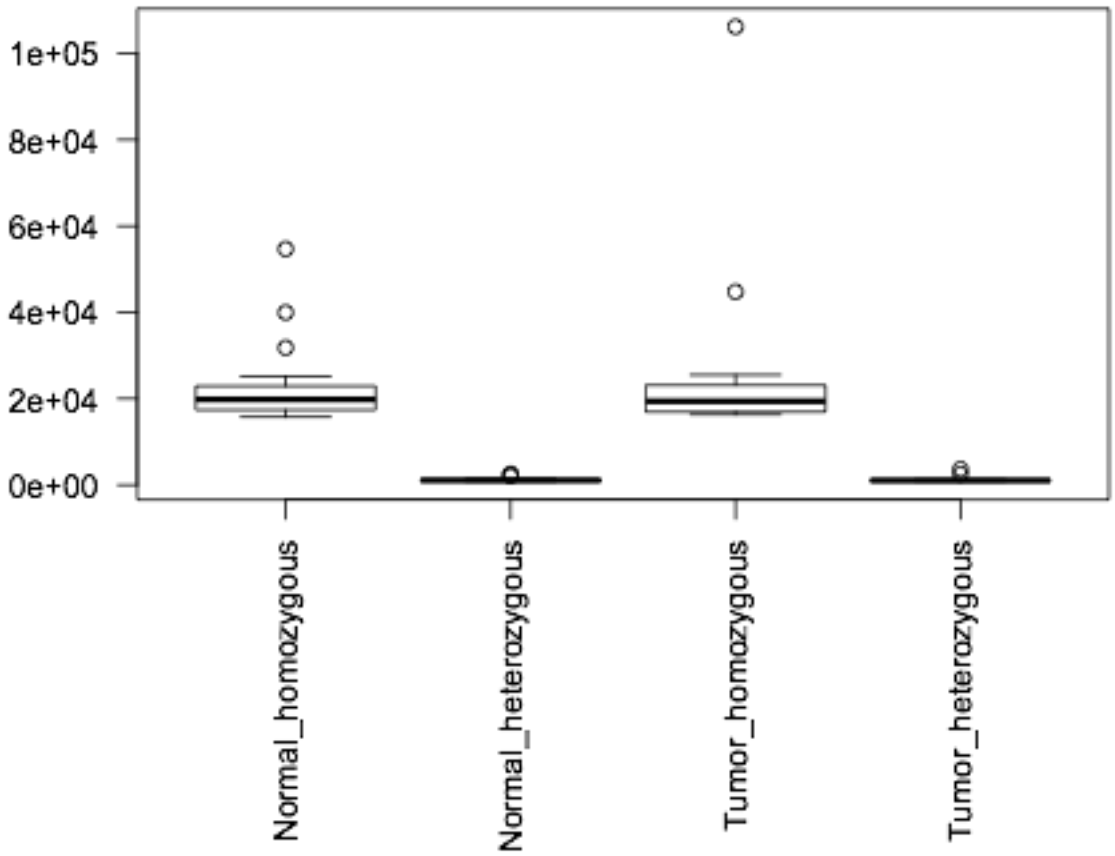
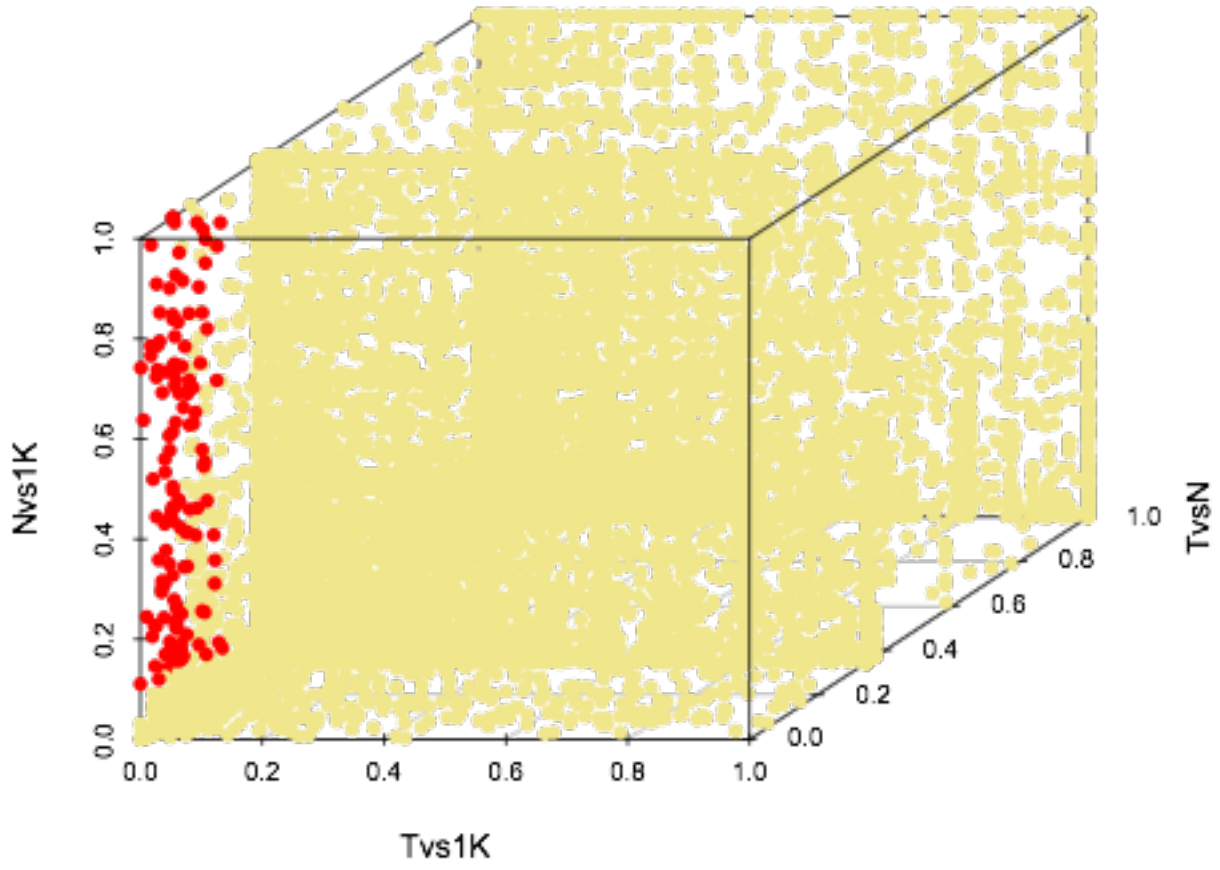


Figure 4.3: Distribution of p-values of all the STRs.

The STRs termed as “cancer causing” are marked in red.



Tables

Table 4.1: Number of concordant, partially discordant and completely discordant sites in the 32 genotyped individuals.

Individual ID	Concordant Sites	Partially Discordant Sites	Completely Discordant Sites
TCGA-22-4609	13981	1116	345
TCGA-22-5471	10953	820	447
TCGA-22-5472	11251	819	450
TCGA-22-5481	10262	853	255
TCGA-22-5482	10695	809	461
TCGA-22-5483	12422	965	346
TCGA-22-5489	11445	883	477
TCGA-22-5491	25574	1644	2161
TCGA-33-6737	11739	866	348
TCGA-34-7107	11652	941	298
TCGA-43-6143	12003	970	361
TCGA-43-6647	12281	907	309
TCGA-43-6771	12073	908	315
TCGA-43-6773	10742	942	256
TCGA-43-7658	9796	797	207
TCGA-51-4081	14942	1433	1415
TCGA-56-8623	11954	1043	274
TCGA-90-6837	9201	759	220
TCGA-33-4587	10535	908	274
TCGA-39-5040	10757	912	252
TCGA-43-5670	9710	854	210
TCGA-77-7142	9694	811	224

Individual ID	Concordant Sites	Partially Discordant Sites	Completely Discordant Sites
TCGA-56-7222	9850	842	237
TCGA-77-7335	9882	824	206
TCGA-77-7337	9873	800	234
TCGA-56-7579	9733	795	216
TCGA-56-7580	10205	829	220
TCGA-56-7582	9543	841	198
TCGA-43-7657	10366	840	231
TCGA-85-7710	10373	847	246
TCGA-56-7730	10128	881	230
TCGA-56-7731	10050	832	202

Table 4.2: Classification of the candidate significant STRs in two cases

(i) STRs that are cancer causing (ii) STRs that have a propensity for cancer

Tumor vs Normal	< 0.1	> 0.1
Tumor vs 1K	< 0.1	< 0.1
Normal vs 1K	> 0.1	< 0.1
Potential Candidate STR Marker Type	cancer markers for diagnosis	cancer predisposition markers

Table 4.3: 58 candidate cancer causing STRs and the genes that they are present within.

Chromosome	Start	End	Motif	Gene
chr1	115112890	115112919	CA	BCAS2
chr1	**175049095**	**175049120**	TG	**TNN**
chr1	176665097	176665143	TG	PAPPA2
chr1	18122802	18122832	TG	ACTL8
chr1	45063638	45063668	CT	RNF220
chr2	113953112	113953156	GGAT	PSD4
chr2	116202656	116202667	CT	DPP10
chr2	207014748	207014778	TCTA	NDUFS1
chr2	**51017274**	**51017301**	CA	**NRXN1**
chr2	68608266	68608290	AG	PLEK
chr2	92119245	92119275	TTTGT	ANAPC1
chr3	107242066	107242094	GCG	BBX
chr3	125270986	125271005	TA	OSBPL11
chr4	3585768	3585785	AC	LINC00955
chr4	74274865	74274889	CA	ALB
chr4	**81433376**	**81433411**	GT	**C4orf22**
chr5	134154413	134154424	AT	DDX46
chr5	136833956	136833990	CA	SPOCK1
chr5	141523758	141523783	TTG	NDFIP1
chr5	73153390	73153423	AC	ARHGEF28
chr6	108214312	108214360	AC	SEC63
chr6	21193536	21193566	AG	CDKAL1
chr6	**39053943**	**39053981**	AC	**GLP1R**

Chromosome	Start	End	Motif	Gene
chr7	134817820	134817851	CA	AGBL3
chr7	**147891229**	**147891256**	GT	**CNTNAP2**
chr7	**30867935**	**30867965**	TG	**INMT-FAM188B / FAM188B**
chr7	50854966	50854992	AGG	GRB10
chr9	104239494	104239524	AAAC	TMEM246
chr9	**137704606**	**137704652**	GT	**COL5A1**
chr10	117856381	117856420	CA	GFRA1
chr10	13912682	13912701	CCTCC	FRMD4A
chr10	6131412	6131428	TCC	RBM17
chr10	82040808	82040822	AC	MAT1A
chr10	**88776060**	**88776106**	TCC	**AGAP11**
chr11	113951168	113951189	CT	ZBTB16
chr12	70732795	70732830	TG	CNOT2
chr13	**110821845**	**110821873**	AC	**COL4A1**
chr13	76102138	76102169	CA	COMMD6
chr13	94727690	94727704	GT	GPC6
chr14	103440619	103440657	AC	CDC42BPB
chr14	23744797	23744825	ATC	HOMEZ
chr15	42950487	42950516	GT	STARD9
chr15	87474798	87474847	TGC	AGBL1
chr16	**22149613**	**22149635**	TG	**VWA3A**
chr16	7713507	7713534	GT	RBFOX1
chr17	10535406	10535457	GTTTT	MYH3
chr17	27162043	27162084	TG	FAM222B
chr17	**40850464**	**40850499**	TG	**CNTNAP1**
chr17	74283112	74283169	CA	QRICH2

Chromosome	Start	End	Motif	Gene
chr17	74476076	74476088	AT	RHBDF2
chr17	9066465	9066502	AC	NTN1
chr18	77030303	77030333	TG	ATP9B
chr19	23938361	23938407	AC	ZNF681
chr19	45537418	45537444	AAAT	RELB
chr19	5986090	5986118	CTC	LOC100128568
chr22	25855683	25855723	GT	CRYBB2P1
chrX	50165596	50165633	AC	DGKK
chrX	**71364555**	**71364596**	CA	**FLJ44635**

Table 4.4: 32 genes that show association with different pathway and GO terms.

Genes	Pathway Term	GO Term
ACTL8		cell differentiation; cellular developmental process
AGBL1		exopeptidase activity
AGBL3		exopeptidase activity
ARHGEF28		neuron projection guidance; axon guidance; neuron projection development; axon development; neuron development; neuron differentiation; cell projection organization; movement of cell or subcellular component; cell differentiation; cell morphogenesis involved in differentiation; neurogenesis; taxis; chemotaxis; cell development; cell projection morphogenesis; cellular developmental process; cell part morphogenesis
CDC42BPB		movement of cell or subcellular component
CDKAL1		4 iron, 4 sulfur cluster binding
CNOT2		cell differentiation; cell development; cellular developmental process
CNTNAP1	Cell adhesion molecules (CAMs)	neuron projection guidance; axon guidance; neuron projection development; axon development; main axon; neuron development; neuron differentiation; cell projection organization; movement of cell or subcellular component; axon part; cell adhesion; cell differentiation; cell morphogenesis involved in differentiation; neurogenesis; neuromuscular process; taxis; chemotaxis; cell development; cell projection morphogenesis; cellular developmental process; cell part morphogenesis; protein binding, bridging
CNTNAP2	Cell adhesion molecules (CAMs)	learned vocalization behavior or vocal learning; neuron projection development; auditory behavior; mechanosensory behavior; main axon; neuron development; vocalization behavior; neuron differentiation; cell projection organization; axon part; response to auditory stimulus; cell adhesion; cell differentiation; neurogenesis; cell development; cellular developmental process; intraspecies interaction between organisms; social behavior

Genes	Pathway Term	GO Term
COL4A1	ECM-receptor interaction - Homo sapiens (human); NCAM1 interactions; Binding and Uptake of Ligands by Scavenger Receptors; Syndecan-1-mediated signaling events; Extracellular matrix organization; miRNA targets in ECM and membrane receptors;	neuron projection guidance; axon guidance; neuron projection development; axon development; platelet-derived growth factor binding; proteinaceous extracellular matrix; neuron development; neuron differentiation; cell projection organization; movement of cell or subcellular component; complex of collagen trimers; cell differentiation; basement membrane; cell morphogenesis involved in differentiation; neurogenesis; taxis; chemotaxis; growth factor binding; cell development; cell projection morphogenesis; cellular developmental process; cell part morphogenesis
COL5A1	ECM-receptor interaction - Homo sapiens (human); Syndecan-1-mediated signaling events; Extracellular matrix organization; miRNA targets in ECM and membrane receptors;	neuron projection guidance; axon guidance; neuron projection development; axon development; platelet-derived growth factor binding; proteinaceous extracellular matrix; neuron development; neuron differentiation; cell projection organization; movement of cell or subcellular component; cell adhesion; complex of collagen trimers; cell differentiation; basement membrane; cell morphogenesis involved in differentiation; neurogenesis; proteoglycan binding; taxis; chemotaxis; growth factor binding; cell development; cell projection morphogenesis; cellular developmental process; cell part morphogenesis
DPP10		exopeptidase activity
FRMD4A		protein binding, bridging
GFRA1	NCAM1 interactions; Signaling events regulated by Ret tyrosine kinase	neuron projection guidance; axon guidance; neuron projection development; axon development; neuron development; neuron differentiation; cell projection organization; movement of cell or subcellular component; cell differentiation; cell morphogenesis involved in differentiation; neurogenesis; taxis; chemotaxis; cell development; cell projection morphogenesis; cellular developmental process; cell part morphogenesis
GPC6		proteinaceous extracellular matrix; movement of cell or subcellular component; proteoglycan binding
GRB10		protein binding, bridging
MYH3		movement of cell or subcellular component; cell differentiation; cell development; cellular developmental process
NDFIP1		cell adhesion; cell differentiation; cellular developmental process
NDUFS1		4 iron, 4 sulfur cluster binding

Genes	Pathway Term	GO Term
NRXN1	Extracellular matrix organization; Cell adhesion molecules (CAMs) - Homo sapiens (human)	learned vocalization behavior or vocal learning; neuron projection guidance; axon guidance; neuron projection development; auditory behavior; axon development; mechanosensory behavior; neuron development; vocalization behavior; neuron differentiation; cell projection organization; movement of cell or subcellular component; axon part; response to auditory stimulus; cell adhesion; cell differentiation; cell morphogenesis involved in differentiation; neurogenesis; neuromuscular process; taxis; chemotaxis; cell development; cell projection morphogenesis; cellular developmental process; cell part morphogenesis; intraspecies interaction between organisms; social behavior
NTN1		neuron projection guidance; axon guidance; neuron projection development; axon development; proteinaceous extracellular matrix; neuron development; neuron differentiation; cell projection organization; movement of cell or subcellular component; cell adhesion; cell differentiation; basement membrane; cell morphogenesis involved in differentiation; neurogenesis; taxis; chemotaxis; cell development; cell projection morphogenesis; cellular developmental process; cell part morphogenesis
OSBPL11		cell differentiation; cellular developmental process
PAPPA2		cell differentiation; cellular developmental process
PLEK		cell projection organization; cell adhesion; cell differentiation; cellular developmental process
PSD4		neuron differentiation; cell differentiation; neurogenesis; cellular developmental process
RBFOX1		neuromuscular process
RELB		cell adhesion; cell differentiation; cellular developmental process
RHBDF2		growth factor binding
SPOCK1		neuron projection development; proteinaceous extracellular matrix; main axon; neuron development; neuron differentiation; cell projection organization; movement of cell or subcellular component; axon part; cell adhesion; cell differentiation; neurogenesis; cell development; cellular developmental process
STARD9		movement of cell or subcellular component

Genes	Pathway Term	GO Term
TNN	ECM-receptor interaction - Homo sapiens (human); Extracellular matrix organization	neuron projection development; axon development; proteinaceous extracellular matrix; neuron development; neuron differentiation; cell projection organization; movement of cell or subcellular component; cell adhesion; cell differentiation; cell morphogenesis involved in differentiation; neurogenesis; cell development; cell projection morphogenesis; cellular developmental process; cell part morphogenesis
ZBTB16		cell adhesion; cell differentiation; cell development; cellular developmental process;

Supplementary Tables

Table S4.1: Clinical data of the 32 Caucasian individuals used in this study.

Sample Name	Years To Birth	Neoplasm Disease Stage	Gender	Number Pack Years Smoked
TCGA-22-4609	81	stage ia	male	77
TCGA-22-5471	75	stage ib	male	50
TCGA-22-5472	67	stage ib	male	35
TCGA-22-5481	72	stage iib	female	75
TCGA-22-5482	81	stage ib	male	60
TCGA-22-5483	74	stage iia	male	50
TCGA-22-5489	64	stage ia	male	70
TCGA-22-5491	74	stage ia	male	57
TCGA-33-6737	71	stage iiia	male	20
TCGA-34-7107	70	stage ii	male	NA
TCGA-43-6143	70	stage ib	male	114
TCGA-43-6647	69	stage iib	female	36
TCGA-43-6771	85	stage ib	male	40
TCGA-43-6773	76	stage iib	male	NA
TCGA-43-7658	75	stage ia	female	12
TCGA-51-4081	55	stage iib	male	80
TCGA-56-8623	71	stage ib	male	45
TCGA-90-6837	64	stage iib	male	40
TCGA-33-4587	63	stage ib	female	NA
TCGA-39-5040	59	stage iiia	male	60
TCGA-43-5670	70	stage iia	male	30

Sample Name	Years To Birth	Neoplasm Disease Stage	Gender	Number Pack Years Smoked
TCGA-77-7142	59	stage ib	female	52.5
TCGA-56-7222	60	NA	male	23
TCGA-77-7335	62	stage iiib	female	52
TCGA-77-7337	65	stage iib	male	79.5
TCGA-56-7579	61	stage iiia	male	2
TCGA-56-7580	84	stage ib	male	NA
TCGA-56-7582	83	stage ib	male	NA
TCGA-43-7657	68	stage ia	female	45
TCGA-85-7710	59	stage ia	female	45
TCGA-56-7730	73	stage iia	male	NA
TCGA-56-7731	66	stage ib	female	100

Table S4.2: 45 candidate cancer causing STRs that are not present within genes.

Chromosome	Start	End	Motif
chr1	16492211	16492235	CA
chr2	122067665	122067695	GT
chr2	165708206	165708242	AC
chr2	187860221	187860258	GT
chr2	681562	681593	TG
chr2	73168265	73168280	TC
chr3	107664808	107664839	GT
chr3	144660787	144660810	GT
chr3	144673554	144673587	TG
chr3	161836228	161836260	CA
chr3	180061458	180061483	TTG
chr3	186911599	186911636	AC
chr3	42022857	42022871	TC
chr3	52341945	52341980	TG
chr3	86140186	86140223	TG
chr3	95684815	95684839	TG
chr4	185899147	185899172	CA
chr4	190976310	190976354	TG
chr4	64522855	64522892	AC
chr5	122835772	122835804	TG
chr5	133126970	133126997	CA
chr5	68529506	68529524	AAAC
chr6	109656194	109656221	AC
chr6	120041747	120041779	AC

Chromosome	Start	End	Motif
chr8	8464373	8464416	TGC
chr8	95256433	95256451	CA
chr9	65628483	65628521	TCCT
chr10	61192049	61192068	GT
chr11	27505324	27505360	AAC
chr11	48248756	48248795	AC
chr12	116202656	116202667	CA
chr12	117061952	117061969	AGG
chr12	96945264	96945302	TG
chr12	97633523	97633538	AT
chr13	24553412	24553447	GT
chr13	64392146	64392187	GT
chr14	32659321	32659356	AC
chr14	98096255	98096288	AC
chr17	41753064	41753100	GT
chr17	41817094	41817129	CA
chr18	53915135	53915166	GT
chr19	7540412	7540446	AC
chr22	48542882	48542918	AC
chrX	112095722	112095756	TG
chrX	147297684	147297712	AC

Table S4.3: List of 624 STRs present within 391 genes that might have a predisposition to lung cancer.

Inclusive of DisGeNet analysis and functional annotation analysis from ConsensusPathDB.

Chromosome	Start	End	Gene
chr10	100229860	100229883	HPSE2
chr10	100250020	100250049	HPSE2
chr10	100481316	100481341	HPSE2
chr10	100773215	100773245	HPSE2
chr10	101977635	101977652	CHUK
chr10	112017617	112017642	MXI1
chr10	112044843	112044867	MXI1
chr10	114911036	114911047	TCF7L2
chr10	123310746	123310761	FGFR2
chr10	126714531	126714545	CTBP2
chr10	127529757	127529769	DHX32
chr10	35338429	35338448	CUL2
chr10	43583481	43583513	RET
chr10	44880457	44880480	CXCL12
chr10	61574358	61574381	CCDC6
chr10	61666055	61666089	CCDC6
chr10	67709916	67709936	CTNNA3
chr10	67934926	67934956	CTNNA3
chr10	68084077	68084104	CTNNA3
chr10	70179227	70179244	DNA2
chr10	75601769	75601784	CAMK2G
chr10	75672213	75672224	PLAU
chr10	90708289	90708315	STAMBPL1

Chromosome	Start	End	Gene
chr10	97604477	97604494	ENTPD1
chr10	97604497	97604516	ENTPD1
chr1	100818008	100818022	CDC14A
chr1	101431175	101431194	SLC30A7
chr1	10292360	10292372	KIF1B
chr1	10318353	10318376	KIF1B
chr1	10318735	10318753	KIF1B
chr1	10363065	10363089	KIF1B
chr1	10510450	10510462	CORT
chr11	110301826	110301861	FDX1
chr11	111783926	111783969	HSPB2
chr11	111808354	111808373	DIXDC1
chr11	114071964	114071990	ZBTB16
chr11	118889386	118889411	TRAPPC4
chr11	118892695	118892709	TRAPPC4
chr11	119077233	119077254	CBL
chr11	119146576	119146599	CBL
chr11	120351230	120351249	ARHGEF12
chr11	121475050	121475077	SORL1
chr11	128358822	128358850	ETS1
chr11	128391627	128391668	ETS1
chr11	18359984	18360004	GTF2H1
chr11	2158447	2158477	IGF2
chr11	27397942	27397963	LGR4
chr1	1276921	1276947	DVL1
chr11	47267478	47267491	ACP2
chr11	47380277	47380291	SPI1
chr1	150790558	150790575	ARNT

Chromosome	Start	End	Gene
chr11	533400	533417	HRAS
chr11	534399	534424	HRAS
chr1	154133366	154133388	TPM3
chr1	156845493	156845528	NTRK1
chr11	58378427	58378449	ZFP91
chr11	58379664	58379681	ZFP91
chr11	58384544	58384572	ZFP91
chr11	62475632	62475649	GNG3
chr1	163297388	163297411	NUF2
chr11	64002386	64002401	VEGFB
chr11	64004558	64004595	VEGFB
chr1	165378730	165378758	RXRG
chr1	165406457	165406471	RXRG
chr11	70269218	70269236	CTTN
chr11	74645033	74645078	XRRA1
chr1	183101428	183101456	LAMC1
chr1	184787663	184787688	FAM129A
chr1	186283934	186283948	TPR
chr1	186295324	186295346	TPR
chr1	186305505	186305525	TPR
chr1	186644332	186644354	PTGS2
chr1	186645847	186645870	PTGS2
chr1	186646573	186646597	PTGS2
chr1	202123247	202123267	PTPN7
chr1	202457853	202457870	PPP1R12B
chr1	206716884	206716921	RASSF5
chr12	104332202	104332221	HSP90B1
chr12	109052795	109052825	CORO1C

Chromosome	Start	End	Gene
chr12	112891293	112891323	PTPN11
chr12	113345089	113345108	OAS1
chr1	211543721	211543754	TRAF5
chr12	117662431	117662442	NOS1
chr12	117680904	117680915	NOS1
chr12	117768106	117768134	NOS1
chr12	120657920	120657936	PXN
chr12	120741280	120741307	SIRT4
chr12	120750091	120750110	SIRT4
chr12	121660877	121660895	P2RX4
chr12	122359386	122359417	WDR66
chr12	12303660	12303677	BCL2L14
chr12	12308129	12308152	BCL2L14
chr1	212482357	212482370	PPP2R5A
chr12	1323832	1323868	ERC1
chr12	1524011	1524037	ERC1
chr12	1553693	1553717	ERC1
chr12	15556755	15556770	PTPRO
chr12	15702203	15702218	PTPRO
chr12	1745869	1745899	WNT5B
chr1	218578686	218578717	TGFB2
chr1	218578727	218578744	TGFB2
chr1	218587570	218587604	TGFB2
chr1	218614738	218614759	TGFB2
chr1	22195990	22196027	HSPG2
chr12	26637009	26637022	ITPR2
chr12	26666682	26666711	ITPR2
chr12	26684588	26684616	ITPR2

Chromosome	Start	End	Gene
chr12	26867982	26867997	ITPR2
chr12	26894605	26894634	ITPR2
chr1	231522983	231523002	EGLN1
chr1	235715207	235715237	GNG4
chr1	241663903	241663940	FH
chr12	4488361	4488401	FGF23
chr12	48465639	48465651	SENP1
chr1	2487988	2488012	TNFRSF14
chr12	51100595	51100611	DIP2B
chr12	54062801	54062822	ATP5G2
chr12	56484484	56484504	ERBB3
chr12	56811590	56811603	TIMELESS
chr12	56814281	56814295	TIMELESS
chr12	56818901	56818927	TIMELESS
chr12	6737610	6737631	LPAR5
chr12	83211337	83211351	TMTC2
chr12	83346290	83346318	TMTC2
chr12	83509562	83509586	TMTC2
chr12	91571920	91571948	DCN
chr13	102433487	102433502	FGF14
chr13	110821801	110821817	COL4A1
chr13	110853576	110853587	COL4A1
chr13	111091158	111091171	COL4A2
chr13	111141587	111141605	COL4A2
chr13	111158523	111158537	COL4A2
chr13	111158654	111158670	COL4A2
chr13	21179094	21179113	IFT88
chr13	22245935	22245947	FGF9

Chromosome	Start	End	Gene
chr13	22246452	22246472	FGF9
chr13	22274151	22274188	FGF9
chr13	28602227	28602248	FLT3
chr13	43537278	43537292	EPSTI1
chr13	52726523	52726534	NEK3
chr13	78475115	78475131	EDNRB
chr14	102551161	102551179	HSP90AA1
chr14	103059389	103059415	RCOR1
chr14	103255932	103255961	TRAF3
chr14	103278066	103278099	TRAF3
chr14	103363423	103363437	TRAF3
chr14	103372472	103372489	TRAF3
chr14	20836924	20836952	TEP1
chr14	23281983	23282001	SLC7A7
chr14	24803757	24803775	ADCY4
chr14	30047243	30047257	PRKD1
chr14	30093606	30093618	PRKD1
chr14	52394577	52394602	GNG2
chr14	54416680	54416718	BMP4
chr14	54423501	54423520	BMP4
chr14	59748243	59748276	DAAM1
chr14	59831895	59831906	DAAM1
chr14	59834006	59834020	DAAM1
chr14	62211579	62211612	HIF1A
chr14	64754820	64754832	ESR2
chr14	67840560	67840591	EIF2S1
chr14	75506463	75506490	MLH3
chr14	75508266	75508282	MLH3

Chromosome	Start	End	Gene
chr14	88442950	88442972	GALC
chr15	100252710	100252743	MEF2A
chr15	26914747	26914771	GABRB3
chr15	27018188	27018210	GABRB3
chr15	27131163	27131185	GABRB3
chr15	34163233	34163246	AVEN
chr15	39886278	39886296	THBS1
chr15	40585566	40585578	PLCB2
chr15	50899297	50899308	TRPM7
chr15	50926524	50926550	TRPM7
chr15	52415087	52415114	GNB5
chr15	57212229	57212251	TCF12
chr15	57539351	57539365	TCF12
chr15	57548742	57548762	TCF12
chr15	64687560	64687572	TRIP4
chr15	64693149	64693178	TRIP4
chr15	64710308	64710330	TRIP4
chr15	67483368	67483392	SMAD3
chr15	73416366	73416383	NEO1
chr15	73492865	73492882	NEO1
chr15	73580491	73580514	NEO1
chr15	79229633	79229647	CTSH
chr15	91030641	91030662	IQGAP1
chr15	91043193	91043221	IQGAP1
chr1	59248124	59248139	JUN
chr15	99251362	99251386	IGF1R
chr16	10733852	10733871	TEKT5
chr16	15832380	15832401	MYH11

Chromosome	Start	End	Gene
chr16	15853171	15853185	MYH11
chr16	15854059	15854087	MYH11
chr16	15894221	15894240	MYH11
chr16	24124909	24124940	PRKCB
chr16	24564880	24564896	RBBP6
chr16	24583566	24583586	RBBP6
chr16	3778401	3778454	CREBBP
chr16	3827669	3827689	CREBBP
chr16	50315500	50315512	ADCY7
chr1	6529183	6529235	PLEKHG5
chr16	55515822	55515844	MMP2
chr16	55517814	55517830	MMP2
chr16	55534407	55534451	MMP2
chr16	56226639	56226663	GNAO1
chr16	56303211	56303233	GNAO1
chr16	56531894	56531907	BBS2
chr16	56535514	56535528	BBS2
chr16	71682675	71682708	PHLPP2
chr16	74682981	74683001	RFWD3
chr16	85743879	85743897	C16orf74
chr16	85944960	85944981	IRF8
chr16	85954339	85954351	IRF8
chr16	89851117	89851143	FANCA
chr1	71438805	71438838	PTGER3
chr17	15968691	15968706	NCOR1
chr17	16012312	16012324	NCOR1
chr17	16021137	16021152	NCOR1
chr17	16256469	16256492	CENPV

Chromosome	Start	End	Gene
chr17	26101215	26101232	NOS2
chr17	27071301	27071319	TRAF4
chr17	30881205	30881238	MYO1D
chr17	31099660	31099686	MYO1D
chr17	31178017	31178039	MYO1D
chr17	33266108	33266126	CCT6B
chr17	33266137	33266156	CCT6B
chr17	36895948	36895963	PCGF2
chr17	39831398	39831418	JUP
chr17	39845230	39845249	JUP
chr17	39914897	39914912	JUP
chr17	40353538	40353578	STAT5B
chr17	40359506	40359543	STAT5B
chr17	40462797	40462819	STAT5A
chr17	40468995	40469007	STAT3
chr17	40518323	40518335	STAT3
chr17	40834653	40834690	CNTNAP1
chr17	40850553	40850586	CNTNAP1
chr17	40850830	40850873	CNTNAP1
chr17	45367815	45367838	ITGB3
chr17	45387723	45387773	ITGB3
chr17	48148114	48148132	ITGA3
chr17	48632411	48632431	SPATA20
chr17	62501101	62501114	DDX5
chr17	63052715	63052732	GNA13
chr17	63528540	63528557	AXIN2
chr17	63533733	63533751	AXIN2
chr17	63550344	63550387	AXIN2

Chromosome	Start	End	Gene
chr17	64574677	64574693	PRKCA
chr17	65822234	65822256	BPTF
chr17	65822267	65822286	BPTF
chr17	65874798	65874819	BPTF
chr17	65900139	65900161	BPTF
chr17	67147967	67147979	ABCA10
chr17	76219856	76219872	BIRC5
chr17	77706456	77706468	ENPP7
chr17	8792106	8792140	PIK3R5
chr18	13051351	13051388	CEP192
chr18	21289734	21289758	LAMA3
chr18	21402079	21402108	LAMA3
chr18	21422555	21422569	LAMA3
chr18	23646565	23646594	SS18
chr18	29110879	29110897	DSG2
chr18	3005675	3005705	LPIN2
chr18	33716512	33716532	ELP2
chr18	39542408	39542419	PIK3C3
chr18	48723138	48723163	MEX3C
chr18	49867015	49867046	DCC
chr18	49868637	49868671	DCC
chr18	49902854	49902888	DCC
chr18	50273337	50273358	DCC
chr18	50561531	50561554	DCC
chr18	50586548	50586574	DCC
chr18	60795697	60795726	BCL2
chr18	6994672	6994698	LAMA1
chr18	7042094	7042105	LAMA1

Chromosome	Start	End	Gene
chr18	7049295	7049313	LAMA1
chr18	7055112	7055138	LAMA1
chr1	9098944	9098959	SLC2A5
chr1	9106725	9106765	SLC2A5
chr19	11129474	11129489	SMARCA4
chr19	11145716	11145730	SMARCA4
chr19	18191906	18191923	IL12RB1
chr19	18284687	18284702	PIK3R2
chr19	2514950	2515003	GNG7
chr19	39664276	39664293	PAK4
chr19	40761346	40761360	AKT2
chr19	41286436	41286463	RAB4B
chr19	42409603	42409642	ARHGEF1
chr19	42420570	42420583	ARHGEF1
chr19	42422077	42422090	ARHGEF1
chr19	44160410	44160432	PLAUR
chr19	45301996	45302018	CBLC
chr19	47249352	47249373	STRN4
chr19	54406558	54406582	PRKCG
chr19	54627013	54627053	PRPF31
chr19	54629853	54629867	PRPF31
chr19	6919417	6919439	EMR1
chr1	9770691	9770713	PIK3CD
chr1	9770715	9770771	PIK3CD
chr1	9770836	9770859	PIK3CD
chr20	16316472	16316492	KIF16B
chr20	16362790	16362816	KIF16B
chr20	16439393	16439440	KIF16B

Chromosome	Start	End	Gene
chr20	33169496	33169508	PIGU
chr20	42340070	42340081	MYBL2
chr20	43610455	43610467	STK4
chr20	43646844	43646870	STK4
chr20	43654960	43654990	STK4
chr20	4835037	4835062	SLC23A2
chr20	60904422	60904447	LAMA5
chr20	6759759	6759778	BMP2
chr20	8721967	8721989	PLCB1
chr20	9274137	9274176	PLCB4
chr20	9360474	9360488	PLCB4
chr20	9449394	9449406	PLCB4
chr2	105704232	105704250	MRPS9
chr2	105706569	105706583	MRPS9
chr2	113594168	113594207	IL1B
chr2	114004578	114004593	PAX8
chr2	114018456	114018497	PAX8
chr2	121656973	121657023	GLI2
chr2	121721014	121721031	GLI2
chr2	121732356	121732379	GLI2
chr21	32513357	32513379	TIAM1
chr21	32575431	32575453	TIAM1
chr21	32623958	32623974	TIAM1
chr21	32652348	32652363	TIAM1
chr21	36382912	36382949	RUNX1
chr21	36840265	36840304	RUNX1
chr21	36972002	36972019	RUNX1
chr21	37204035	37204052	RUNX1

Chromosome	Start	End	Gene
chr21	38310970	38310989	HLCS
chr21	44316955	44316977	NDUFV3
chr21	47678652	47678672	MCM3AP
chr2	15742815	15742840	DDX1
chr2	174086202	174086218	ZAK
chr2	17912865	17912882	SMC6
chr2	17959406	17959421	SMC6
chr2	17959442	17959455	SMC6
chr2	187521715	187521751	ITGAV
chr2	202151326	202151353	CASP8
chr2	20921309	20921327	C2orf43
chr2	21000842	21000868	C2orf43
chr2	212488818	212488849	ERBB4
chr2	212951888	212951918	ERBB4
chr2	216262044	216262063	FN1
chr2	216269021	216269050	FN1
chr2	216274854	216274875	FN1
chr2	216287904	216287942	FN1
chr2	217329414	217329426	SMARCAL1
chr2	219724790	219724807	WNT6
chr2	219756335	219756353	WNT10A
chr22	21298693	21298711	CRKL
chr22	21307713	21307740	CRKL
chr22	23605340	23605356	BCR
chr22	28939411	28939447	TTC28
chr22	28943183	28943223	TTC28
chr22	30234193	30234217	ASCC2
chr22	33218895	33218917	TIMP3

Chromosome	Start	End	Gene
chr22	37326236	37326252	CSF2RB
chr22	37334728	37334750	CSF2RB
chr22	39636900	39636914	PDGFB
chr22	39636900	39636919	PDGFB
chr22	39746187	39746206	SYNGR1
chr2	242287625	242287647	SEPT2
chr2	242289469	242289481	SEPT2
chr2	33759335	33759351	RASGRP3
chr2	33764108	33764124	RASGRP3
chr2	33774434	33774450	RASGRP3
chr2	39233852	39233863	SOS1
chr2	42996710	42996757	HAAO
chr2	46605202	46605217	EPAS1
chr2	47797390	47797408	MSH2
chr2	48018282	48018315	MSH6
chr2	79423059	79423077	CTNNA2
chr2	79516412	79516432	CTNNA2
chr2	79548920	79548945	CTNNA2
chr2	79851109	79851124	CTNNA2
chr2	79999148	79999169	CTNNA2
chr2	80004454	80004502	CTNNA2
chr2	80116303	80116325	CTNNA2
chr2	80458084	80458111	CTNNA2
chr2	80598629	80598649	CTNNA2
chr3	100357855	100357901	GPR128
chr3	121500480	121500510	IQCB1
chr3	121500516	121500528	IQCB1
chr3	123102629	123102657	ADCY5

Chromosome	Start	End	Gene
chr3	124523877	124523920	ITGB5
chr3	124592720	124592738	ITGB5
chr3	12650636	12650649	RAF1
chr3	128342834	128342866	RPN1
chr3	134264298	134264313	CEP63
chr3	134265194	134265223	CEP63
chr3	138116077	138116091	MRAS
chr3	138116379	138116413	MRAS
chr3	138433626	138433644	PIK3CB
chr3	138454800	138454816	PIK3CB
chr3	13874869	13874881	WNT7A
chr3	13914507	13914531	WNT7A
chr3	141239385	141239402	RASA2
chr3	141299398	141299426	RASA2
chr3	158296278	158296293	MLF1
chr3	158296295	158296318	MLF1
chr3	158409078	158409100	GFM1
chr3	169234899	169234917	MECOM
chr3	169352025	169352057	MECOM
chr3	171336639	171336668	PLD1
chr3	171444127	171444145	PLD1
chr3	171975360	171975399	FNDC3B
chr3	185199102	185199120	MAP3K13
chr3	191984584	191984618	FGF12
chr3	191988576	191988628	FGF12
chr3	192126850	192126890	FGF12
chr3	193130400	193130429	ATP13A4
chr3	25611473	25611512	RARB

Chromosome	Start	End	Gene
chr3	41268677	41268690	CTNNB1
chr3	42251578	42251609	TRAK1
chr3	45167744	45167760	CDCP1
chr3	4562834	4562858	ITPR1
chr3	4572458	4572479	ITPR1
chr3	4715262	4715285	ITPR1
chr3	47913667	47913689	MAP4
chr3	48019805	48019821	MAP4
chr3	48092583	48092617	MAP4
chr3	4857670	4857700	ITPR1
chr3	58120319	58120334	FLNB
chr3	60650174	60650201	FHIT
chr3	69834418	69834445	MITF
chr3	8775809	8775833	CAV3
chr3	99787195	99787227	FILIP1L
chr4	103504124	103504138	NFKB1
chr4	103611952	103611974	MANBA
chr4	107248408	107248430	AIMP1
chr4	108985791	108985808	LEF1
chr4	110897463	110897475	EGF
chr4	110929469	110929484	EGF
chr4	113970781	113970795	ANK2
chr4	113970784	113970795	ANK2
chr4	113998153	113998203	ANK2
chr4	114214679	114214693	ANK2
chr4	114469974	114469992	CAMK2D
chr4	120548075	120548101	PDE5A
chr4	120549755	120549800	PDE5A

Chromosome	Start	End	Gene
chr4	154623897	154623913	TLR2
chr4	164467100	164467136	MARCH1
chr4	164511210	164511234	MARCH1
chr4	164637369	164637391	MARCH1
chr4	165118286	165118311	MARCH1
chr4	165183117	165183136	MARCH1
chr4	177713373	177713389	VEGFC
chr4	1809111	1809146	FGFR3
chr4	187118035	187118055	CYP4V2
chr4	187179067	187179079	KLKB1
chr4	39184298	39184315	WDR19
chr4	53773876	53773910	SCFD2
chr4	53826877	53826917	SCFD2
chr4	54043605	54043627	SCFD2
chr4	54255949	54255971	PDGFRA
chr4	54319248	54319261	PDGFRA
chr4	54588112	54588129	PDGFRA
chr4	54678366	54678398	PDGFRA
chr4	54940757	54940798	PDGFRA
chr4	55069013	55069032	PDGFRA
chr4	55980135	55980181	KDR
chr4	55981749	55981779	KDR
chr4	76862043	76862060	NAAA
chr4	77532395	77532424	SHROOM3
chr4	77589728	77589754	SHROOM3
chr4	8219844	8219863	SH3TC1
chr4	8224404	8224423	SH3TC1
chr4	8235285	8235305	SH3TC1

Chromosome	Start	End	Gene
chr4	84228562	84228583	HPSE
chr4	86574045	86574063	ARHGAP24
chr4	86661360	86661374	ARHGAP24
chr4	86863152	86863168	ARHGAP24
chr4	87422834	87422856	MAPK10
chr4	88903020	88903056	SPP1
chr4	9982937	9982993	SLC2A9
chr5	125929001	125929036	ALDH7A1
chr5	131539876	131539894	P4HA2
chr5	131599917	131599947	PDLIM4
chr5	142022862	142022882	FGF1
chr5	145522675	145522689	LARS
chr5	146966978	146967014	JAKMIP2
chr5	147003392	147003415	JAKMIP2
chr5	147016689	147016702	JAKMIP2
chr5	147063860	147063897	JAKMIP2
chr5	149492850	149492880	CSF1R
chr5	149501689	149501708	PDGFRB
chr5	149610968	149611007	CAMK2A
chr5	149619103	149619122	CAMK2A
chr5	149624902	149624944	CAMK2A
chr5	150487713	150487735	ANXA6
chr5	150506273	150506294	ANXA6
chr5	156525921	156525942	HAVCR2
chr5	31418316	31418347	DROSHA
chr5	35081875	35081916	PRLR
chr5	35207717	35207745	PRLR
chr5	44388715	44388732	FGF10

Chromosome	Start	End	Gene
chr5	60229766	60229784	ERCC8
chr5	7489928	7489944	ADCY2
chr5	7525769	7525793	ADCY2
chr5	76673817	76673834	PDE8B
chr5	77311119	77311131	AP3B1
chr5	77563270	77563283	AP3B1
chr5	96103737	96103755	ERAP1
chr6	105298957	105298976	HACE1
chr6	106975082	106975095	AIM1
chr6	112450373	112450412	LAMA4
chr6	112462465	112462486	LAMA4
chr6	112528148	112528175	LAMA4
chr6	112528562	112528573	LAMA4
chr6	114292110	114292134	HDAC2
chr6	126075583	126075606	HEY2
chr6	129274743	129274770	LAMA2
chr6	129465358	129465378	LAMA2
chr6	129663352	129663375	LAMA2
chr6	129692393	129692425	LAMA2
chr6	129712930	129712974	LAMA2
chr6	129826640	129826670	LAMA2
chr6	133072135	133072165	VNN2
chr6	152129546	152129569	ESR1
chr6	152382090	152382108	ESR1
chr6	159208242	159208254	EZR
chr6	20424606	20424641	E2F3
chr6	20483277	20483300	E2F3
chr6	29545632	29545656	GABBR1

Chromosome	Start	End	Gene
chr6	29643880	29643902	ZFP57
chr6	29694337	29694360	HLA-F
chr6	29899155	29899176	HLA-G
chr6	29968735	29968761	HLA-G
chr6	3017255	3017312	NQO2
chr6	31543874	31543886	TNF
chr6	33168087	33168115	RXRB
chr6	33173749	33173775	HSD17B8
chr6	33623668	33623679	ITPR3
chr6	33625593	33625618	ITPR3
chr6	36034519	36034558	MAPK14
chr6	36043546	36043564	MAPK14
chr6	36075047	36075077	MAPK14
chr6	43749877	43749897	VEGFA
chr6	49421247	49421259	MUT
chr6	56226553	56226592	COL21A1
chr6	7378685	7378700	CAGE1
chr7	100274423	100274435	GNB2
chr7	106522521	106522536	PIK3CG
chr7	107564543	107564555	LAMB1
chr7	111103233	111103254	IMMP2L
chr7	111136942	111136996	IMMP2L
chr7	116166859	116166876	CAV1
chr7	116932342	116932372	WNT2
chr7	121021711	121021745	FAM3C
chr7	122764733	122764762	SLC13A1
chr7	123324547	123324570	WASL
chr7	126393442	126393454	GRM8

Chromosome	Start	End	Gene
chr7	127314183	127314211	SND1
chr7	127344766	127344785	SND1
chr7	128490154	128490168	FLNC
chr7	128844846	128844871	SMO
chr7	140573361	140573372	BRAF
chr7	150155222	150155253	GIMAP8
chr7	150438172	150438200	GIMAP5
chr7	21640073	21640096	DNAH11
chr7	21713647	21713680	DNAH11
chr7	39663232	39663270	RALA
chr7	42069787	42069805	GLI3
chr7	44286587	44286604	CAMK2B
chr7	45699359	45699371	ADCY1
chr7	55135785	55135815	EGFR
chr7	75242404	75242425	HIP1
chr7	75256198	75256231	HIP1
chr8	101206162	101206174	SPAG1
chr8	120865125	120865142	DSCC1
chr8	131859520	131859555	ADCY8
chr8	17400699	17400741	SLC7A2
chr8	38285914	38285933	FGFR1
chr8	41548427	41548446	ANK1
chr8	41566636	41566654	ANK1
chr8	41574744	41574763	ANK1
chr8	41615862	41615878	ANK1
chr8	42177058	42177081	IKBKB
chr8	42186868	42186879	IKBKB
chr8	76470718	76470741	HNF4G

Chromosome	Start	End	Gene
chr8	92972371	92972385	RUNX1T1
chr8	95952409	95952426	TP53INP1
chr8	95952674	95952698	TP53INP1
chr9	113774750	113774782	LPAR1
chr9	124461708	124461725	DAB2IP
chr9	127975853	127975866	RABEPK
chr9	128509664	128509685	PBX3
chr9	128677932	128677954	PBX3
chr9	133701811	133701826	ABL1
chr9	133759202	133759224	ABL1
chr9	133947528	133947542	LAMC3
chr9	136565838	136565875	SARDH
chr9	136629252	136629277	VAV2
chr9	136657150	136657162	VAV2
chr9	136667399	136667412	VAV2
chr9	136719145	136719157	VAV2
chr9	137231342	137231379	RXRA
chr9	137249777	137249798	RXRA
chr9	140069365	140069388	ANAPC2
chr9	140080911	140080940	ANAPC2
chr9	33138969	33138987	B4GALT1
chr9	33150967	33150990	B4GALT1
chr9	34342853	34342877	NUDT2
chr9	74840905	74840950	GDA
chr9	79438778	79438793	PRUNE2
chr9	97584253	97584294	C9orf3
chr9	97588986	97589004	C9orf3
chr9	98214611	98214639	PTCH1

Chromosome	Start	End	Gene
chr9	98270188	98270215	PTCH1
chr9	98278959	98278974	PTCH1
chrX	107397098	107397120	COL4A6
chrX	107423675	107423709	COL4A6
chrX	107818014	107818033	COL4A5
chrX	107840190	107840224	COL4A5
chrX	123026664	123026688	XIAP
chrX	133057159	133057179	GPC3
chrX	133119384	133119398	GPC3
chrX	153596673	153596692	FLNA
chrX	47496205	47496222	ELK1
chrX	66943337	66943361	AR
chrX	66943363	66943382	AR

Table S4.4: Pathway terms associated with the 58 genes that have short tandem repeats with variation in LUSC tumor samples when compared to normal and 1000 Genomes population data.

Pathway	Members Input Overlap	P-value
ECM-receptor interaction - Homo sapiens (human)	COL5A1; TNN; COL4A1	0.002
NCAM1 interactions	GFRA1; COL4A1	0.0052
Signaling events regulated by Ret tyrosine kinase	GRB10; GFRA1	0.0058
Binding and Uptake of Ligands by Scavenger Receptors	ALB; COL4A1	0.0061
Spliceosome - Homo sapiens (human)	RBM17; BCAS2; DDX46	0.0068
Syndecan-1-mediated signaling events	COL5A1; COL4A1	0.007
Extracellular matrix organization	COL5A1; TNN; NRXN1; COL4A1	0.0071
miRNA targets in ECM and membrane receptors	COL5A1; COL4A1	0.0077
Cell adhesion molecules (CAMs) - Homo sapiens (human)	CNTNAP1; CNTNAP2; NRXN1	0.0082
Vitamin B12 Metabolism	ALB; MAT1A	0.0098

Table S4.5: GO terms associated with the 58 genes that have short tandem repeats with variation in LUSC tumor samples when compared to normal and 1000 Genomes population data.

term_name	term_goid	members_input_overlap_geneids	p-value
learned vocalization behavior or vocal learning	GO:0098598	CNTNAP2; NRXN1	0.0001
neuron projection guidance	GO:0097485	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; COL4A1	0.0001
axon guidance	GO:0007411	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; COL4A1	0.0001
neuron projection development	GO:0031175	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; SPOCK1; COL4A1; TNN; CNTNAP2	0.0002
auditory behavior	GO:0031223	CNTNAP2; NRXN1	0.0002
axon development	GO:0061564	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; COL4A1; TNN	0.0003
platelet-derived growth factor binding	GO:0048407	COL5A1; COL4A1	0.0004
mechanosensory behavior	GO:0007638	CNTNAP2; NRXN1	0.0005
proteinaceous extracellular matrix	GO:0005578	COL4A1; NTN1; COL5A1; SPOCK1; GPC6; TNN	0.0005
main axon	GO:0044304	CNTNAP1; CNTNAP2; SPOCK1	0.0005
neuron development	GO:0048666	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; SPOCK1; COL4A1; TNN; CNTNAP2	0.0007
vocalization behavior	GO:0071625	CNTNAP2; NRXN1	0.0007
neuron differentiation	GO:0030182	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; SPOCK1; COL4A1; PSD4; TNN; CNTNAP2	0.001

term_name	term_goid	members_input_overlap_geneids	p-value
cell projection organization	GO:0030030	CNTNAP2; NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; SPOCK1; COL4A1; TNN; PLEK	0.001
movement of cell or subcellular component	GO:0006928	MYH3; STARD9; CDC42BPB; NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; SPOCK1; COL4A1; GPC6; TNN	0.001
axon part	GO:0033267	CNTNAP1; CNTNAP2; SPOCK1; NRXN1	0.001
response to auditory stimulus	GO:0010996	CNTNAP2; NRXN1	0.001
cell adhesion	GO:0007155	CNTNAP2; ZBTB16; NTN1; NRXN1; CNTNAP1; NDFIP1; COL5A1; SPOCK1; RELB; TNN; PLEK	0.002
complex of collagen trimers	GO:0098644	COL5A1; COL4A1	0.002
cell differentiation	GO:0030154	MYH3; CNTNAP2; ZBTB16; OSBPL11; NTN1; PAPPA2; PSD4; ACTL8; NRXN1; CNTNAP1; ARHGEF28; NDFIP1; COL5A1; GFRA1; SPOCK1; COL4A1; RELB; CNOT2; TNN; PLEK	0.002
basement membrane	GO:0005604	NTN1; COL5A1; COL4A1	0.002
cell morphogenesis involved in differentiation	GO:0000904	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; COL4A1; TNN	0.003
neurogenesis	GO:0022008	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; SPOCK1; COL4A1; PSD4; TNN; CNTNAP2	0.003
proteoglycan binding	GO:0043394	COL5A1; GPC6	0.0033
exopeptidase activity	GO:0008238	DPP10; AGBL3; AGBL1	0.0034
neuromuscular process	GO:0050905	CNTNAP1; RBFOX1; NRXN1	0.0036
taxis	GO:0042330	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; COL4A1	0.0039

term_name	term_goid	members_input_overlap_geneids	p-value
chemotaxis	GO:0006935	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; COL4A1	0.0039
growth factor binding	GO:0019838	RHBDF2; COL5A1; COL4A1	0.0047
cell development	GO:0048468	MYH3; ZBTB16; NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; SPOCK1; COL4A1; CNOT2; TNN; CNTNAP2	0.0049
cell projection morphogenesis	GO:0048858	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; COL4A1; TNN	0.005
cellular developmental process	GO:0048869	MYH3; CNTNAP2; ZBTB16; OSBPL11; NTN1; PAPPA2; PSD4; ACTL8; NRXN1; CNTNAP1; ARHGEF28; NDFIP1; COL5A1; GFRA1; SPOCK1; COL4A1; RELB; CNOT2; TNN; PLEK	0.0053
cell part morphogenesis	GO:0032990	NTN1; NRXN1; CNTNAP1; ARHGEF28; COL5A1; GFRA1; COL4A1; TNN	0.0057
4 iron, 4 sulfur cluster binding	GO:0051539	NDUFS1; CDKAL1	0.0065
intraspecies interaction between organisms	GO:0051703	CNTNAP2; NRXN1	0.0078
social behavior	GO:0035176	CNTNAP2; NRXN1	0.0078
protein binding, bridging	GO:0030674	CNTNAP1; GRB10; FRMD4A	0.0098

Chapter Five

Conclusion and Future Directions

This thesis outlines initial studies in the field of improving somatic variant identification from multiple platforms and identification of short tandem repeats as biomarkers for cancer. We developed the first framework to improve the identification of somatic variants on whole genome and exome platforms and identified STRs as new biomarkers that could help the prognosis of lung squamous cell carcinoma. I further discuss the challenges and the tremendous amount of work to be done in the field of somatic variant identification and identification of new biomarkers for cancer.

5.1 Understanding of somatic variants

My first work in this thesis aims to understand the identification of somatic variant callers in different scenarios. Here, we compared 12 pipelines encompassing three of the most widely used mappers and four somatic variant callers using simulated and real data sets, which is unprecedented in scale for somatic variant calling.

A well-characterized set of somatic variants of an individual is a need of the hour to help advance the field of somatic variant calling. The Genome in a Bottle Consortium produced a well characterized set of germline variants for the individual, NA12878 from the 1000 Genomes Project. This highly confident set of mutations was developed by integrating 14 whole exome and whole genome data sets that were sequenced on five platforms and whose mutations were called using seven mappers and three variant callers. Such a well-characterized set of variants of an individual has to be produced for somatic variants as well to remove the biases from the simulated data set.

With the passage of time, the number of somatic variant callers and other methods to identify somatic variants continues to increase. A Web platform like GCAT (www.bioplanet.com/gcat) [87] was developed to compare tools that identify germline variants. GCAT has results from several popular mappers and germline variant callers as public reports. GCAT helps users make a fair comparison of their newly developed methods with existing pipelines consisting of mappers and variant callers. A similar platform can be developed for somatic variant caller comparison. Although the TCGA-ICGC dream challenge uses crowd sourcing to determine what tools work better to identify somatic variants, users cannot download the data from the TCGA-ICGC platform and check their results immediately. Hence, it is necessary to develop tools similar to GCAT for somatic variant callers.

5.2 Improving identification of somatic variants

We developed the first tool in integrating somatic variant calling from both the whole genome and whole exome platforms to identify somatic variants more accurately. We were able to achieve an average F1 score of 0.77 for simulated data and 0.681 for real data. Although, the F1 score of 0.681 seems low, it is still higher than the F1 score that would have been achieved by using only one somatic variant caller. A systematic analysis needs to be done to show which machine learning algorithm works best in different case scenarios. The simulated data generated in the current work had 700 simulated somatic variants (positives) and 9300 base positions that were not somatic (negatives). One tool in particular, RacedIncrementalLogitBoost, had an F1 score of 0 from a 10-fold cross-

validation using the above simulated data, but an F1 score of 0.839 using the combination of A15K-A0BW-A152 as the training set and A15E as the test set. The reason for the low F1 score was due to the low number of variants present in the training set of the simulated data. As A15K-A0BW-A152 had 2419 variants in the training set, RacedIncrementalLogitBoost performed better in the case of real data. This shows the need for a comprehensive analysis on which machine learning tool can be used to improve identification of somatic variants.

In a clinical setup, it is still uncommon for normal and tumor samples to be both sequenced due to the cost. Since MuTect has a version that can identify somatic variants using only the tumor sample with less sensitivity and VCMM uses only tumor samples, this work could be extended to identifying somatic variants from only tumor samples using MuTect and VCMM. We used both whole genome and whole exome platforms to identify somatic variants more accurately. If sequencing cost reduces further and sequencing quality improves more, we would be in this interesting phase of precision medicine where whole genome sequencing will take place intermittently during one's lifetime and we should be able to identify variants, both germline and somatic, accurately from one platform.

5.3 Identification of short tandem repeats as new biomarkers for cancer

It is important to understand the entire spectrum of variants that are present in cancer to gain insight into which variants are driver mutations and which variants are passenger mutations. This would help us discern the mutations that occur during the progression of tumor. With this in mind, we studied the variation of 285,063 short tandem repeats in squamous cell lung cancer and identified the STRs that can serve as cancer marker. An extension of this study would be to identify potential STR markers for other types of cancer. Table 5.1 shows the number of people whose exome and mRNA have been sequenced for different kinds of cancer in TCGA. Another aspect of this analysis is to do a whole genome analysis to study the trans-effect of short tandem repeats on the genes. The analysis should reveal the same and different short tandem repeats that vary in different cancers. We can then further study if the variation in short tandem repeats are driver mutations or passenger mutations. This could help us gain insight into different short tandem repeats that are affected in different cancers.

5.4 Conclusion

This thesis improves our understanding of which somatic variant callers work better under which conditions for the identification of somatic variants. We developed a new framework for more accurate identification of somatic variants, when two different platforms give different results. We also look into short tandem repeats as potential biomarkers for cancer. We also specify the need to understand the entire mutation

spectrum in the human genome to help improve the diagnosis, prognosis, and therapy of cancer.

Tables

Table 5.1: Number of individuals (at least 10) whose matched normal-tumor exome and mRNA have been sequenced for different kinds of cancer in TCGA

Number of samples	Cancer Type
71	Kidney Renal Clear Cell Carcinoma
70	Breast Invasive Cancer
58	Thyroid Carcinoma
58	Lung Adenocarcinoma
48	Lung Hepatocellular Carcinoma
44	Lung Squamous Cell Carcinoma
43	Prostate Adenocarcinoma
39	Head and Neck Squamous Cell Carcinoma
37	Colon Adenocarcinoma
32	Kidney Renal Papillary Cell Carcinoma
30	Stomach Adenocarcinoma
25	Kidney Chromophobe
21	Uterine Corpus Endometrioid Carcinoma
19	Bladder Urothelial Carcinoma
13	Esophageal Carcinoma

Bibliography

1. Lander ES, Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al.: Initial sequencing and analysis of the human genome. Nature 2001, 409:860-921.

2. International Human Genome Sequencing C: Finishing the euchromatic sequence of the human genome. Nature 2004, 931-945.

3. Wetterstrand K: DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).

4. Mardis ER: Next-Generation DNA Sequencing Methods. Annual Review of Genomics and Human Genetics 2008, 9:387-402.

5. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M et al: An integrated semiconductor device enabling non-optical genome sequencing. Nature 2011, 475(7356):348-352.

6. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B et al: Real-time DNA sequencing from single polymerase molecules. Science (New York, NY) 2009, 323(5910):133-138.

7. Cheryl Heiner PB, Susana Wang, Yan Guo, Meredith Ashby, Joan Wilson, Kevin, Travers JC, and Jason Underwood: Greater than 10 kb Read Lengths Routine when Sequencing with Pacific Biosciences' XL Release. http://wwwpacificbiosciencescom/pdf/Poster_PAG2013_GreaterThan10kbReadspdf 2013.

8. Hayden EC: Is the $1,000 genome for real? http://wwwnaturecom/news/is-the-1-000-genome-for-real-114530 2014.

9. https://www.genome.gov/27541954/dna-sequencing-costs/.

10. Feuk L, Carson AR, Scherer SW: Structural variation in the human genome. Nature reviews Genetics 2006, 7(2):85-97.

11. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ et al: Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet 2016, 48(1):22-29.

12. Gonzaga-Jauregui C, Lupski JR, Gibbs RA: Human genome sequencing in health and disease. Annual review of medicine 2012, 63:35-61.

13. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA et al: Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. The New England journal of medicine 2010, 362(13):1181-1191.

14. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M et al: Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science (New York, NY) 2010, 328(5978): 636-639.

15. McMahon Francis J, Insel Thomas R: Pharmacogenomics and Personalized Medicine in Neuropsychiatry. Neuron 2012, 74(5):773-776.

16. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS et al: International network of cancer genome projects. Nature 2010, 464(7291):993-998.

17. Ding L, Wendl MC, Koboldt DC, Mardis ER: Analysis of next-generation genomic data in cancer: accomplishments and challenges. Human molecular genetics 2010, 19(R2):R188-196.

18. Matthew N. Bainbridge WW, David R. Murdock, Jennifer Friedman, Claudia Gonzaga-Jauregui, Irene Newsham, Jeffrey G. Reid, John K. Fink, Margaret B. Morgan, Marie-Claude Gingras, Donna M. Muzny, Linh D. Hoang, Shahed Yousaf, James R. Lupski, and Richard A. Gibbs: Whole-Genome Sequencing for Optimized Patient Management. Science Translational Medicine 2011, 3:87re83.

19. Kalow W: Pharmacogenetics and pharmacogenomics: origin, status, and the hope for personalized medicine. The pharmacogenomics journal 2006, 6(3):162-165.

20. Olson MV VA: Sequencing the chimpanzee genome: Insights into human evolution and disease. Nature Reviews Genetics 2003, 4(1):20-28.

21. Weber-Lehmann J, Schilling E, Gradl G, Richter DC, Wiehler J, Rolf B: Finding the needle in the haystack: differentiating "identical" twins in paternity testing and forensics by ultra-deep next generation sequencing. Forensic science international Genetics 2014, 9:42-46.

22. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: The human genome browser at UCSC. Genome Res 2002, 12(6):996-1006.

23. The Genomes Project C: A global reference for human genetic variation. Nature 2015, 526(7571):68-74.

24. Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, O'Donnell-Luria A, Ware J, Hill A, Cummings B et al: Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv 2015.

25. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC et al: A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet 2015, 47(10):1121-1130.

26. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ et al: Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science (New York, NY) 2007, 316(5829):1331-1336.

27. http://www.autismconsortium.org/

28. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001, 29(1): 308-311.

29. Bauer-Mehren A, Rautschka M, Sanz F, Furlong LI: DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. Bioinformatics 2010, 26(22):2924-2926.

30. Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000, 28(1):27-30.

31. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B et al: International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database : the journal of biological databases and curation 2011, 2011:bar026.

32. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR: The Catalogue of Somatic Mutations in Cancer (COSMIC). Current protocols in human genetics / editorial board, Jonathan L Haines [et al] 2008, Chapter 10:Unit 10.11.

33. Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, Francke U, Mountain JL, Goldman SM, Tanner CM, Langston JW et al: Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS genetics 2011, 7(6):e1002141.

34. Fonseca NA, Rung, J., Brazma, A., and Marioni, J.C.: Tools for mapping high-throughput sequencing data. Bioinformatics 2012, 28:3169-3177.

35. http://en.wikipedia.org/wiki/List_of_sequence_alignment_software

36. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. Nat Methods 2012, 9(4):357-359.

37. http://www.novocraft.com

38. TMAP: the Torrent Mapping Alignment Program. [https://github.com/iontorrent/TMAP]

39. David M, Dzamba M, Lister D, Ilie L, Brudno M: SHRiMP2: sensitive yet practical SHort Read Mapping. Bioinformatics 2011, 27(7):1011-1012.

40. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M: SHRiMP: accurate mapping of short color-space reads. PLoS computational biology 2009, 5(5):e1000386.

41. Lunter G, Goodson M: Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 2011, 21(6):936-939.

42. Come Raczy RP, Christopher T. Saunders, Ilya Chorny, Semyon Kruglyak, Elliott H. Margulies, Han-Yu Chuang, Morten Källberg, Swathi A. Kumar, Arnold Liao, Kristina M. Little, Michael P. Strömberg and Stephen W. Tanner: Isaac: Ultra-fast whole genome secondary analysis on Illumina sequencing platforms. Bioinformatics 2013, 29(16):2041-2043.

43. Li H, Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing,: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25:2078-2079.

44. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, Chuang HY, Kallberg M, Kumar SA, Liao A et al: Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics 2013, 29(16):2041-2043.

45. McKenna A HM, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 2010, 20(9):1297-1303.

46. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25(16):2078-2079.

47. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C, Milosavljevic A, Gibbs RA, Yu F: An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 2012, 13:8.

48. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012, 22(3): 568-576.

49. Daichi Shigemizu AF, Shintaro Akiyama, Tetsuo Abe, Kaoru Nakano, Keith A. Boroevich, Yujiro Yamamoto, Mayuko Furuta, Michiaki Kubo, Hidewaki Nakagawa, and Tatsuhiko Tsunodaa: A practical method to detect SNVs and indels from whole genome and exome sequencing data. Scientific Reports 2013, 3(2161).

50. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L: SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 2012, 28(3):311-317.

51. Chiang DY, Getz G, Jaffe DB, O'Kelly MJ, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES: High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Methods 2009, 6(1):99-103.

52. Miller CA, Hampton O, Coarfa C, Milosavljevic A: ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PloS one 2011, 6(1):e16327.

53. Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA et al: Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proceedings of the National Academy of Sciences of the United States of America 2011, 108(46):E1128-1136.

54. Mayrhofer M, DiLorenzo S, Isaksson A: Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. Genome Biol 2013, 14(3):R24.

55. Yau C: OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. Bioinformatics 2013, 29(19):2482-2484.

56. http://compbio.bccrc.ca/software/hmmcopy/.

57. http://www.stjuderesearch.org/site/lab/zhang.

58. Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, Conroy JC, Wang J, Liu S: Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. Oncotarget 2013, 4(11):1868-1881.

59. Sathirapongsasuti JF, Lee H, Horst BA, Brunner G, Cochran AJ, Binder S, Quackenbush J, Nelson SF: Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics 2011, 27(19): 2648-2654.

60. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA et al: Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol 2012, 30(5):413-421.

61. Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, Barillot E: Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics 2011, 27(2):268-269.

62. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP et al: BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 2009, 6(9): 677-681.

63. Sun R, Love MI, Zemojtel T, Emde AK, Chung HR, Vingron M, Haas SA: Breakpointer: using local mapping artifacts to support sequence breakpoint discovery from single-end reads. Bioinformatics 2012, 28(7):1024-1025.

64. Marschall T, Costa IG, Canzar S, Bauer M, Klau GW, Schliep A, Schonhuth A: CLEVER: clique-enumerating variant finder. Bioinformatics 2012, 28(22): 2875-2882.

65. Sindi SS, Onal S, Peng LC, Wu HT, Raphael BJ: An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biol 2012, 13(3):R22.

66. Wong K, Keane TM, Stalker J, Adams DJ: Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol 2010, 11(12):R128.

67. Mohiyuddin M, Mu JC, Li J, Bani Asadi N, Gerstein MB, Abyzov A, Wong WH, Lam HY: MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. Bioinformatics 2015, 31(16):2741-2744.

68. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z: A survey of tools for variant analysis of

next-generation genome sequencing data. Briefings in bioinformatics 2014, 15(2): 256-278.

69. Gymrek M, Golan D, Rosset S, Erlich Y: lobSTR: A short tandem repeat profiler for personal genomes. Genome Res 2012, 22(6):1154-1162.

70. Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D: Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Res 2013, 41(1):e32.

71. Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y: The landscape of human STR variation. Genome Res 2014, 24(11):1894-1904.

72. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al: Initial sequencing and analysis of the human genome. Nature 2001, 409(6822):860-921.

73. Mardis ER: Next-generation DNA sequencing methods. Annual review of genomics and human genetics 2008, 9:387-402.

74. Greater than 10 kb Read Lengths Routine when Sequencing with Pacific Biosciences' XL Release [http://wwwpacificbiosciencescom/pdf/ Poster_PAG2013_GreaterThan10kbReadspdf]

75. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D et al: The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 2010, 465(7297):473-477.

76. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin ML, Ordonez GR, Bignell GR et al: A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 2010, 463(7278): 191-196.

77. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, Turashvili G, Ding J, Tse K, Haffari G et al: The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature 2012, 486(7403):395-399.

78. http://www.biostars.org/

79. Li J-W, Robison K, Martin M, Sjödin A, Usadel B, Young M, Olivares EC, Bolser DM: The SEQanswers wiki: a wiki database of tools for high-throughput sequencing analysis. Nucleic Acids Research 2012, 40(D1):D1313-D1317.

80. Kim SY, Speed TP: Comparing somatic mutation-callers: beyond Venn diagrams. BMC Bioinformatics 2013, 14:189.

81. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M: Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotech 2014, 32(3):246-251.

82. Li H, Ruan J, Durbin R: Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008, 18(11):1851-1858.

83. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In: arXiv. vol. 1303.3997; 2013.

84. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013, 31(3):213-219.

85. Shigemizu D, Fujimoto A, Akiyama S, Abe T, Nakano K, Boroevich KA, Yamamoto Y, Furuta M, Kubo M, Nakagawa H et al: A practical method to detect SNVs and indels from whole genome and exome sequencing data. Scientific Reports 2013, 3:2161.

86. Grada A, Weinbrecht K: Next-generation sequencing: methodology and application. The Journal of investigative dermatology 2013, 133(8):e11.

87. Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, Mittelman D: An analytical framework for optimizing variant discovery from personal genomes. Nature communications 2015, 6:6275.

88. Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, Dahlman KB, Pao W, Zhao Z: Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. Genome Med 2013, 5(10):91.

89. Fang H, Wu Y, Narzisi G, O'Rawe JA, Barron LT, Rosenbaum J, Ronemus M, Iossifov I, Schatz MC, Lyon GJ: Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med 2014, 6(10):89.

90. David E. Larson CCH, Ken Chen, Daniel C. Koboldt, Travis E. Abbott, David J. Dooling, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding: SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 2012, 28(3):311-317.

91. Koboldt DC ZQ, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research 2012, 22(3): 568-576.

92. Kristian Cibulskis MSL, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, Gad Getz: Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology 2013, 31:213-219.

93. Vijayan V, Zhang L: Evaluation of pipelines detecting somatic point variants and analysis of factors affecting the detection. Under Review.

94. Adam D Ewing KEH, Yin Hu, Kyle Ellrott, Cristian Caloian, Takafumi N Yamaguchi, J Christopher Bare, Christine P'ng, Daryl Waggott, Veronica Y Sabelnykova, ICGC-TCGA DREAM Somatic Mutation Calling Challenge participants, Michael R Kellen, Thea C Norman, David Haussler, Stephen H Friend, Gustavo Stolovitzky, Adam A Margolin, Joshua M Stuart, Paul C Boutros: Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nature Methods 2015, 12:623-630.

95. Su Yeon Kim LJaTPS: Combining calls from multiple somatic mutation-callers. BMC Bioinformatics 2014, 15(154).

96. Li Tai Fang PTA, Aparna Chhibber, Marghoob Mohiyuddin, Yu Fan, John C. Mu, Greg Gibeling, Sharon Barr, Narges Bani Asadi, Mark B. Gerstein, Daniel C. Koboldt, Wenyi Wang, Wing H. Wong and Hugo YK Lam: An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biology 2015, 16(197).

97. Mamunur Rashid CDR-E, Alistair G. Rust, and David J. Adams: Cake: a bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. 2013 2013, 29(17):2208–2210.

98. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova J-L, Abel L: Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proceedings of the National Academy of Sciences 2015, 112(17):5473-5478.

99. The Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013, 45(10): 1113-1120.

100. Puente XS, Pinyol M, Quesada V, Conde L, Ordonez GR, Villamor N, Escaramis G, Jares P, Bea S, Gonzalez-Diaz M et al: Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature 2011, 475(7354): 101-105.

101. Mamunur Rashid CDR-E, Alistair G. Rust, and David J. Adams: Cake: a bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. Bioinformatics 2013, 29(17):2208-2210.

102. Mark Hall EF, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten: The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009, 11(1).

103. David L Goode SMH, Maria A Doyle, Tao Ma, Simone M Rowley, David Choong, Georgina L Ryland and Ian G Campbell: A simple consensus approach improves somatic mutation prediction accuracy. Genome Medicine 2013, 5(9).

104. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM: Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010, 127(12):2893-2917.

105. Govindan R, Ding L, Griffith M, Subramanian J, Dees Nathan D, Kanchi Krishna L, Maher Christopher A, Fulton R, Fulton L, Wallis J et al: Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. Cell 2012, 150(6):1121-1134.

106. Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA: Non–Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship. Mayo Clinic proceedings Mayo Clinic 2008, 83(5):584-594.

107. Thomas A: Small cell lung cancer in profile. Science translational medicine 2015, 7(299):299ec135-299ec135.

108. Onganer PU, Seckl MJ, Djamgoz MBA: Neuronal characteristics of small-cell lung cancer. British Journal of Cancer 2005, 93(11):1197-1201.

109. Samet JM, Avila-Tang E, Boffetta P, Hannan LM, Olivo-Marston S, Thun MJ, Rudin CM: LUNG CANCER IN NEVER SMOKERS: CLINICAL EPIDEMIOLOGY AND ENVIRONMENTAL RISK FACTORS. Clinical cancer research : an official journal of the American Association for Cancer Research 2009, 15(18):5626-5645.

110. http://www.cancer.org/acs/groups/cid/documents/webcontent/003115-pdf.pdf.

111. Shen C, Wang X, Tian L, Che G: Microsatellite alteration in multiple primary lung cancer. Journal of Thoracic Disease 2014, 6(10):1499-1505.

112. Fondon JW, 3rd, Martin A, Richards S, Gibbs RA, Mittelman D: Analysis of microsatellite variation in Drosophila melanogaster with population-scale genome sequencing. PloS one 2012, 7(3):e33036.

113. Lupski JR: Genomic rearrangements and sporadic disease. Nat Genet 2007.

114. Donald F Conrad JEMK, Mark A DePristo, Sarah J Lindsay, Yujun Zhang, Ferran Casals, Youssef Idaghdour, Chris L Hartl, Carlos Torroja, Kiran V Garimella, Martine Zilversmit, Reed Cartwright, Guy A Rouleau, Mark Daly, Eric A Stone, Matthew E Hurles, Philip Awadalla for the 1000 Genomes Project: Variation in genome-wide mutation rates within and between human families. Nat Genet 2011, 43(7):712-714.

115. Mirkin SM: Expandable DNA repeats and human disease. Nature 2007, 447(7147): 932-940.

116. Clarke LA, Rebelo CS, Goncalves J, Boavida MG, Jordan P: PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. Molecular pathology : MP 2001, 54(5):351-353.

117. Benatti P, Gafà R, Barana D, Marino M, Scarselli A, Pedroni M, Maestri I, Guerzoni L, Roncucci L, Menigatti M et al: Microsatellite Instability and Colorectal Cancer Prognosis. American Association for Cancer Research 2005, 11(23):8332-8340.

118. Boland CR, Goel A: Microsatellite instability in colorectal cancer. Gastroenterology 2010, 138(6):2073-2087.e2073.

119. Hall G, Clarkson A, Shi A, Langford E, Leung H, Eckstein RP, Gill AJ: Immunohistochemistry for PMS2 and MSH6 alone can replace a four antibody panel for mismatch repair deficiency screening in colorectal adenocarcinoma. Pathology 2010, 42(5):409-413.

120. Hirose T, Kondo K, Takahashi Y, Ishikura H, Fujino H, Tsuyuguchi M, Hashimoto M, Yokose T, Mukai K, Kodama T et al: Frequent microsatellite instability in lung cancer from chromate-exposed workers. Molecular carcinogenesis 2002, 33(3): 172-180.

121. Asicioglu F, Oguz-Savran F, Ozbek U: Mutation rate at commonly used forensic STR loci: paternity testing experience. Disease markers 2004, 20(6):313-315.

122. http://cpdb.molgen.mpg.de/

123. B A, A J, J L, al. e: Molecular Biology of the Cell., vol. 4th edition: Garland Science; 2002.

124. http://www.genecards.org/

125. Szatmári T, Dobra K: The Role of Syndecan-1 in Cellular Signaling and its Effects on Heparan Sulfate Biosynthesis in Mesenchymal Tumors. Frontiers in Oncology 2013, 3:310.

126. Crnic I, Strittmatter K, Cavallaro U, Kopfstein L, Jussila L, Alitalo K, Christofori G: Loss of Neural Cell Adhesion Molecule Induces Tumor Metastasis by Up-regulating Lymphangiogenesis. Cancer Research 2004, 64(23):8630-8638.

127. Canton J, Neculai D, Grinstein S: Scavenger receptors in homeostasis and immunity. Nat Rev Immunol 2013, 13(9):621-634.

128. Ben J, Jin G, Zhang Y, Ma B, Bai H, Chen J, Zhang H, Gong Q, Zhou X, Zhang H et al: Class A Scavenger Receptor Deficiency Exacerbates Lung Tumorigenesis by Cultivating a Procarcinogenic Microenvironment in Humans and Mice. American Journal of Respiratory and Critical Care Medicine 2012, 186(8):763-772.

129. Shaw AT, Hsu PP, Awad MM, Engelman JA: Tyrosine kinase gene rearrangements in epithelial malignancies. Nat Rev Cancer 2013, 13(11):772-787.

130. van Alphen RJ, Wiemer EAC, Burger H, Eskens FALM: The spliceosome as target for anticancer treatment. Br J Cancer 2008, 100(2):228-232.

131. Piyathilake CJ, Johanning GL, Macaluso M, Whiteside M, Oelschlager DK, Heimburger DC, Grizzle WE: Localized folate and vitamin B-12 deficiency in squamous cell lung cancer is associated with global DNA hypomethylation. Nutrition and cancer 2000, 37(1):99-107.

132. Muramatsu T, Muramatsu H, Kojima T: Identification of proteoglycan-binding proteins. Methods in enzymology 2006, 416:263-278.

133. Ma Y-Q, Geng J-G: Heparan Sulfate-Like Proteoglycans Mediate Adhesion of Human Malignant Melanoma A375 Cells to P-Selectin Under Flow. The Journal of Immunology 2000, 165(1):558-565.

134. Sekido Y, Takahashi T, Ueda R, Takahashi M, Suzuki H, Nishida K, Tsukamoto T, Hida T, Shimokata K, Zsebo KM et al: Recombinant human stem cell factor mediates chemotaxis of small-cell lung cancer cell lines aberrantly expressing the c-kit protooncogene. Cancer Res 1993, 53(7):1709-1714.

135. Zhang Z, Zhou Y, Qian H, Shao G, Lu X, Chen Q, Sun X, Chen D, Yin R, Zhu H et al: Stemness and inducing differentiation of small cell lung cancer NCI-H446 cells. Cell Death & Disease 2013, 4(5):e633.

136. Molenaar WM, de Leij L, Trojanowski JQ: Neuroectodermal tumors of the peripheral and the central nervous system share neuroendocrine N-CAM-related

antigens with small cell lung carcinomas. Acta neuropathologica 1991, 83(1): 46-54.

137. Chedotal A, Kerjan G, Moreau-Fauvarque C: The brain within the tumor: new roles for axon guidance molecules in cancers. Cell Death Differ 2005, 12(8): 1044-1056.

138. Bai F, Morcos F, Sohn YS, Darash-Yahana M, Rezende CO, Lipper CH, Paddock ML, Song L, Luo Y, Holt SH et al: The Fe-S cluster-containing NEET proteins mitoNEET and NAF-1 as chemotherapeutic targets in breast cancer. Proceedings of the National Academy of Sciences of the United States of America 2015, 112(12):3698-3703.

139. Song SW, Jun BC, Cho KJ, Lee S, Kim YJ, Park SH: CT Evaluation of Vocal Cord Paralysis due to Thoracic Diseases: A 10-Year Retrospective Study. Yonsei Medical Journal 2011, 52(5):831-837.

140. Wang T, Elston RC: Improved power by use of a weighted score test for linkage disequilibrium mapping. American journal of human genetics 2007, 80(2): 353-360.

141. Gauderman WJ, Murcray C, Gilliland F, Conti DV: Testing association between disease and multiple SNPs in a candidate gene. Genetic epidemiology 2007, 31(5):383-395.

142. Buil A, Martinez-Perez A, Perera-Lluna A, Rib L, Caminal P, Soria JM: A new gene-based association test for genome-wide association studies. BMC proceedings 2009, 3 Suppl 7:S130.

143. Wan Y-W, Allen GI, Liu Z: TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. Bioinformatics 2016, 32(6):952-954.

144. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM et al: MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 2010, 38(18):e178.

145. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011, 12(1):1-16.

146. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD, Gould MN, Stewart RM, Kendziorski C: EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 2013, 29(8): 1035-1043.